АНАЛИЗ ИСТОРИИ ПРОДАЖ ОБЪЕКТОВ НЕДВИЖИМОСТИ

Автор:
Леонов В.В.
М23-524

Преподаватель:
Киреев В.С.
к.т.н., в.н.с., доцент

Национальный исследовательский ядерный университет «МИФИ»
г. Москва

Цели и задачи исследования

Провести углубленный анализ истории продаж объектов недвижимости с использованием методов факторного и кластерного анализа для выявления закономерностей

Предобработка данных: очистка набора данных от пропущенных значений (NaN), а также добавление нескольких новых признаков для улучшения качества анализа.

Факторный анализ признаков: использование методов факторного анализа для уменьшения размерности и выявления скрытых факторов, влияющих на цену недвижимости.

Кластеризация данных: применение методов кластерного анализа для выделения групп объектов недвижимости с схожими характеристиками, что позволяет лучше понять структуру рынка.

Визуализация результатов: создание визуальных представлений для лучшего понимания полученных результатов и выявления ключевых трендов

Набор данных

Набор о сделках с недвижимостью США за период с 2001 по 2022 год
1 097 629 x 14

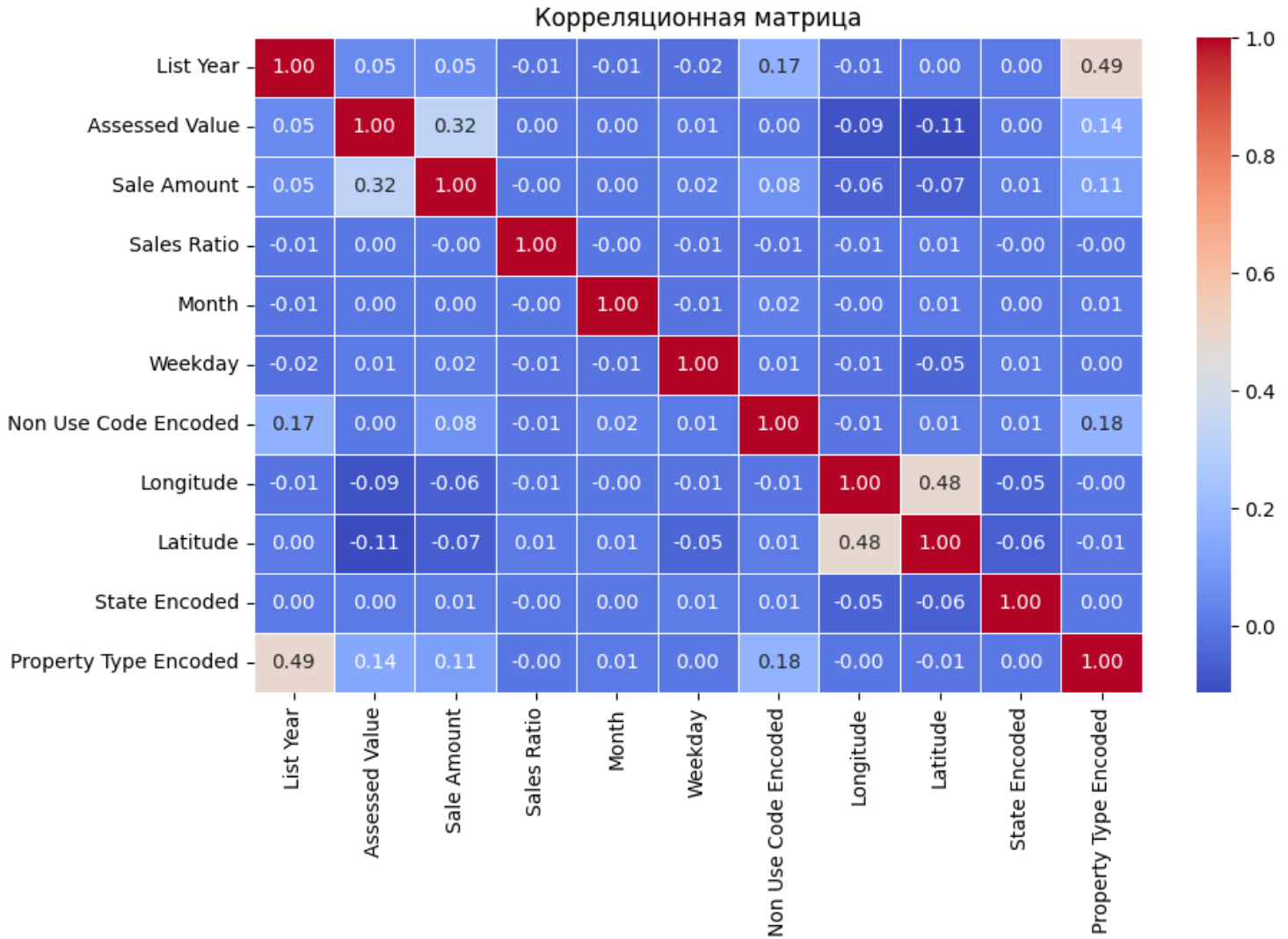
- Удаление столбцов «Residential Type», «Assessor Remarks» и «OPM remarks» и строк с Nan в столбцах «Property Type», «Non Use Code» и «Location»
- Замена категориальные признаки «Property Type» и «Non Use Code» на числовые значения: для этого был использован метод замены категориальных значений на среднее значение «Sale Amount» для каждой категории.
- Извлечение дня недели и месяца из столбца с датой сделки, определение штата по координатам и его замена числовым значением
- Итоговый набор
59042 x 19

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.

Корреляционная матрица Пирсона



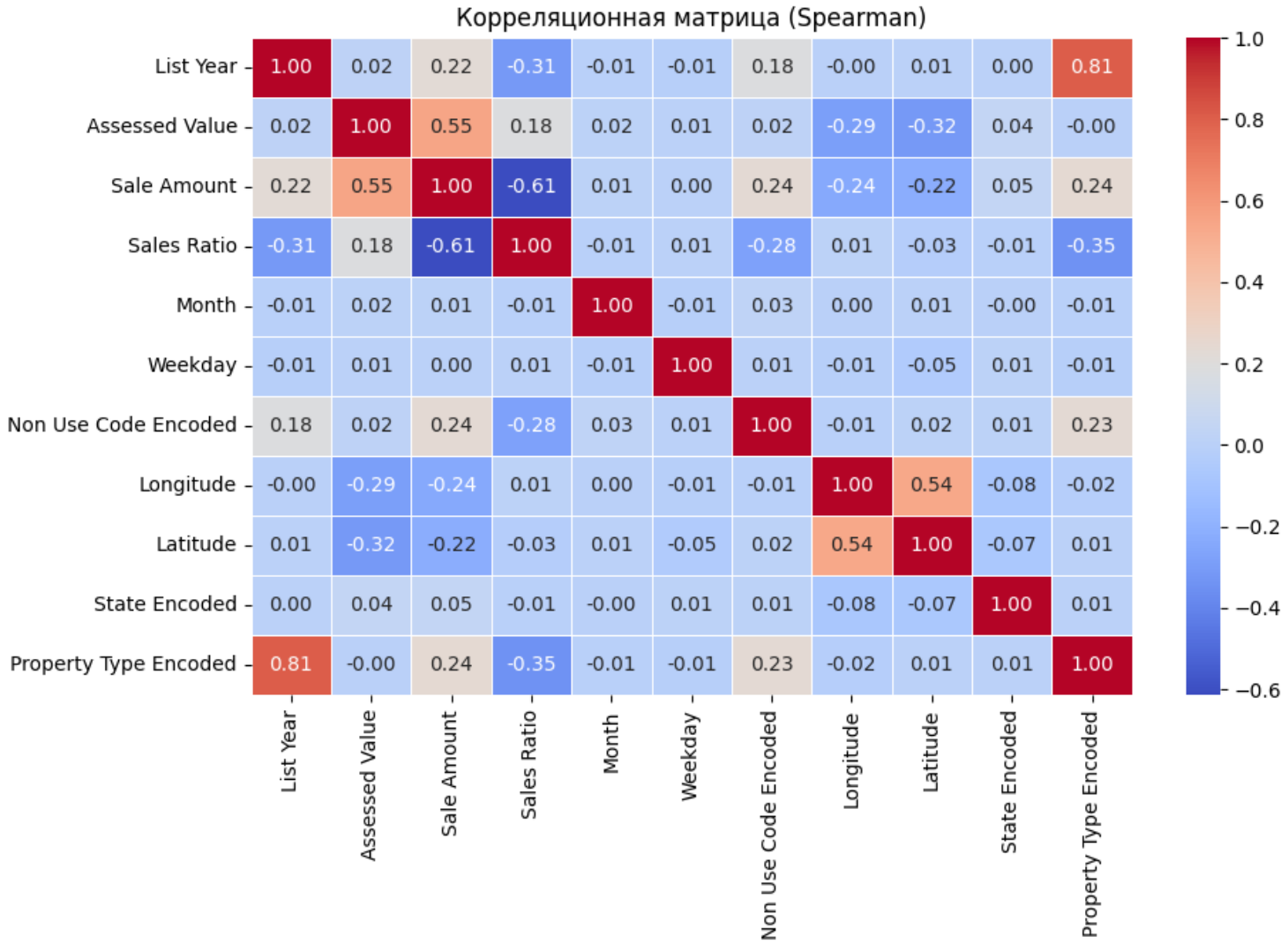
«List Year» и «Property Type Encoded»
(корреляция = 0.49)

«Assessed Value» и «Sale Amount»
(корреляция = 0.32)

«Non Use Code Encoded» и «Property Type Encoded»
(корреляция = 0.18)

«Assessed Value» и «Property Type Encoded»
(корреляция = 0.14)

Корреляционная матрица Спирмана



«List Year» и «Property Type Encoded»
(корреляция = 0.81)

«Assessed Value» и «Sale Amount»
(корреляция = 0.55)

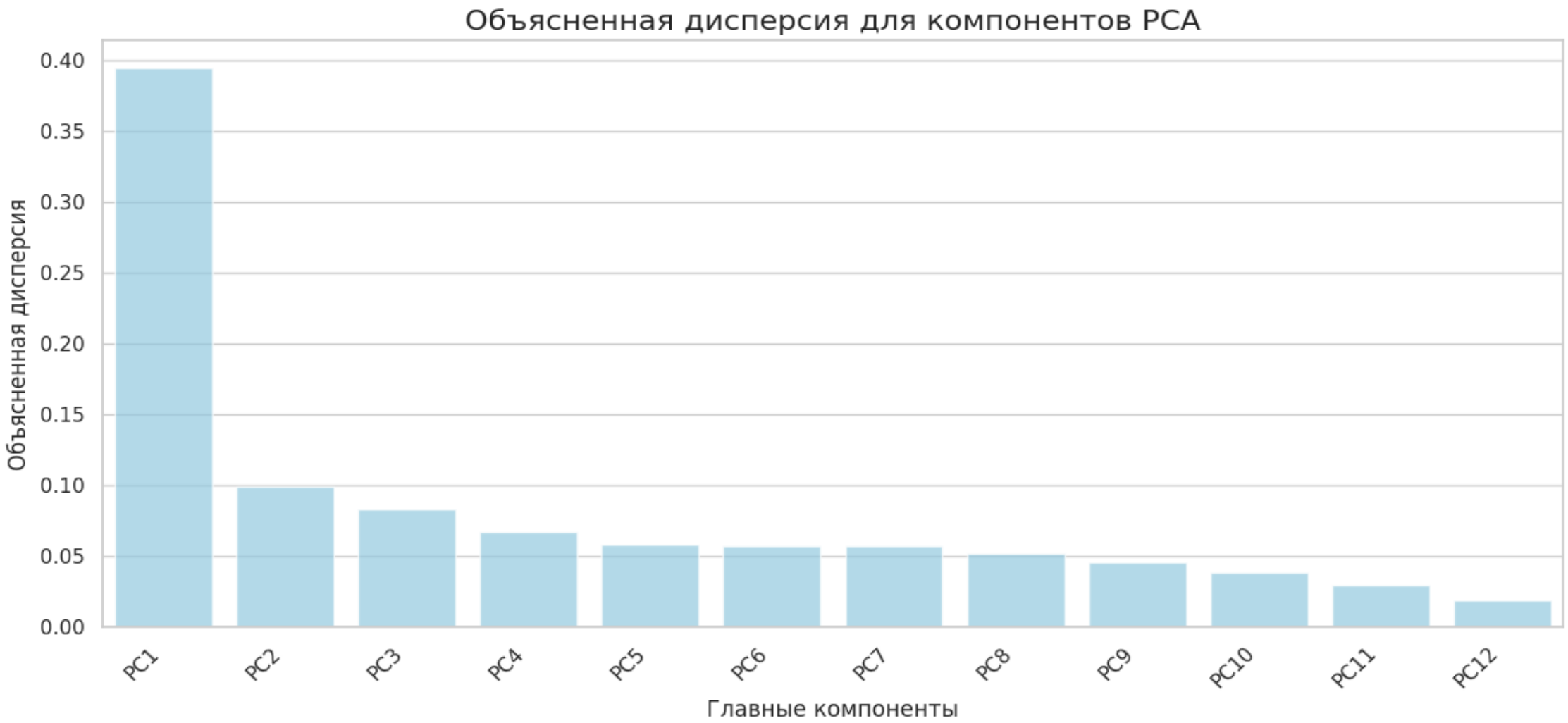
«Longitude» и «Latitude»
(корреляция = 0.54)

«Sales Ratio» и «Sale Amount»
(корреляция = -0.61)

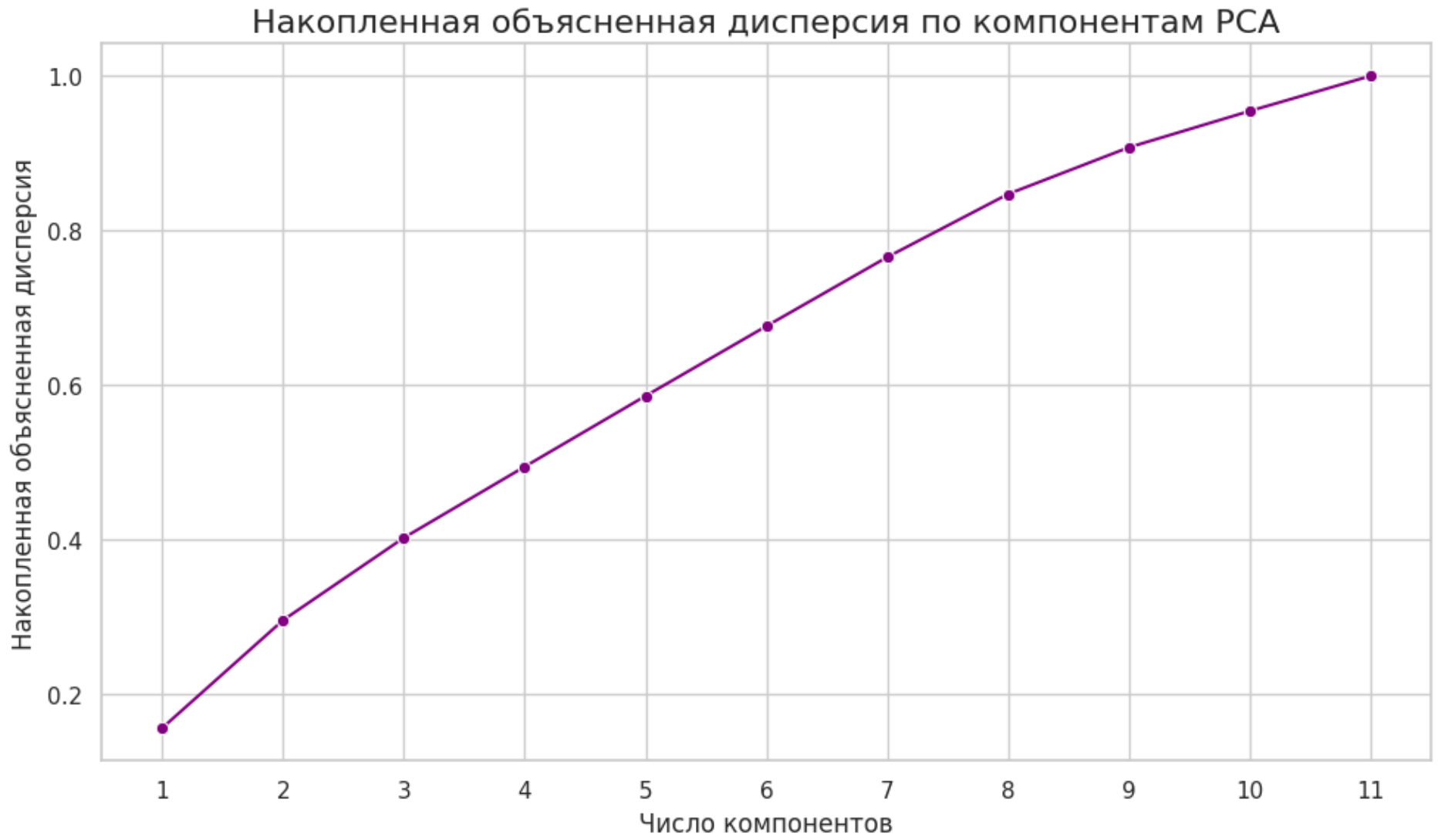
«Sales Ratio» и «Property Type Encoded»
(корреляция = -0.35)

«List Year» и «Sales Ratio»
(корреляция = -0.31)

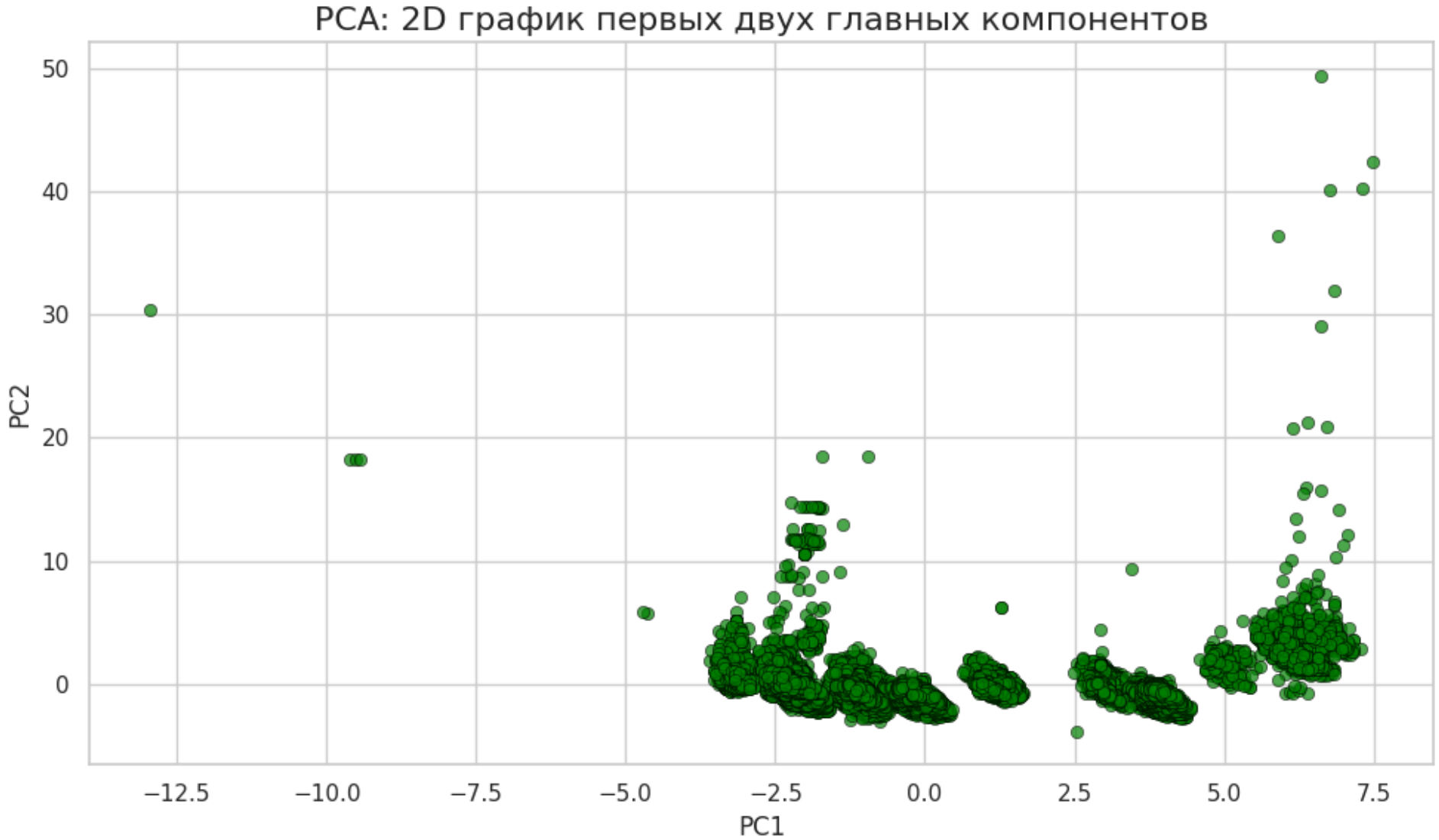
Анализ главных компонент



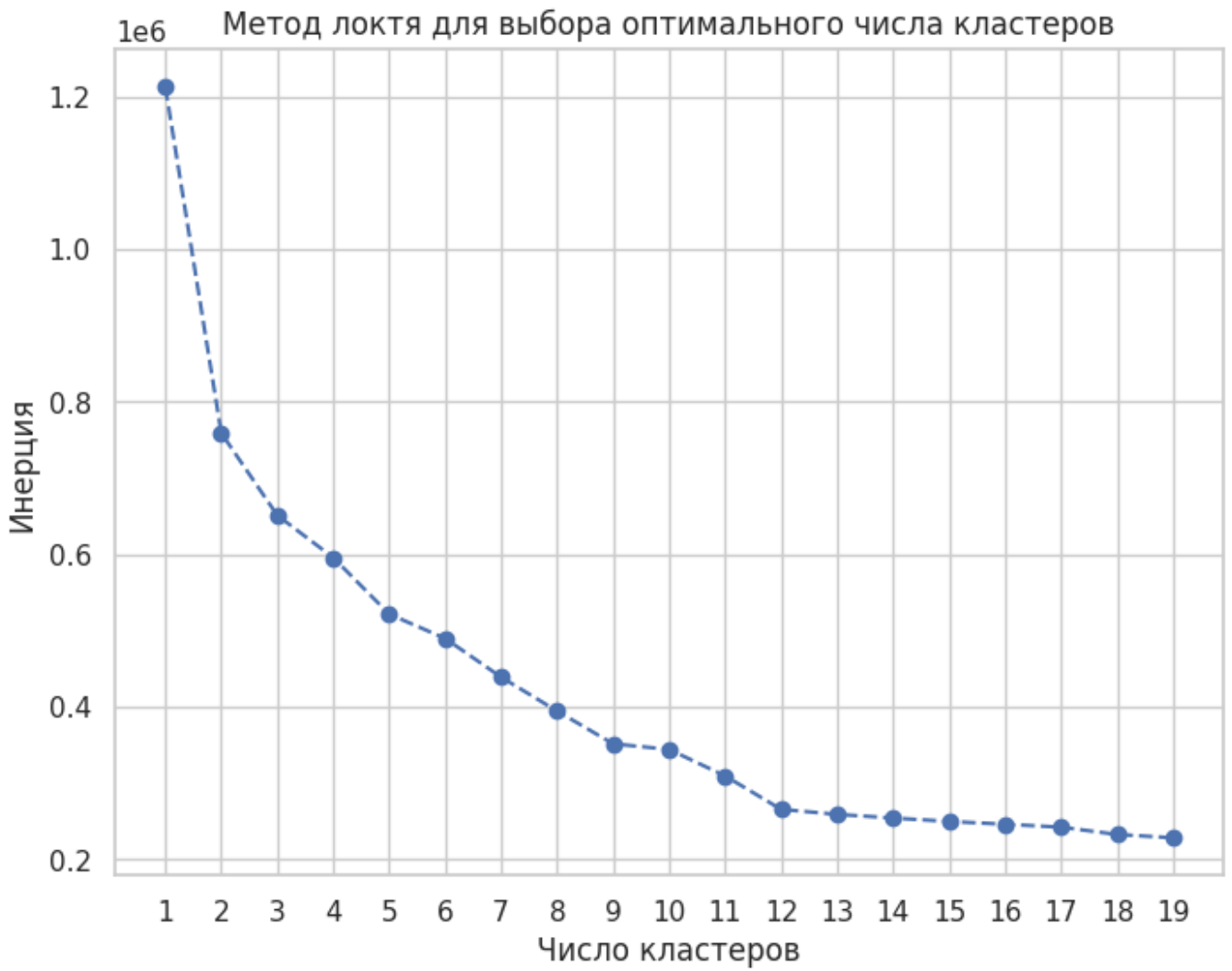
Анализ главных компонент



Анализ главных компонент



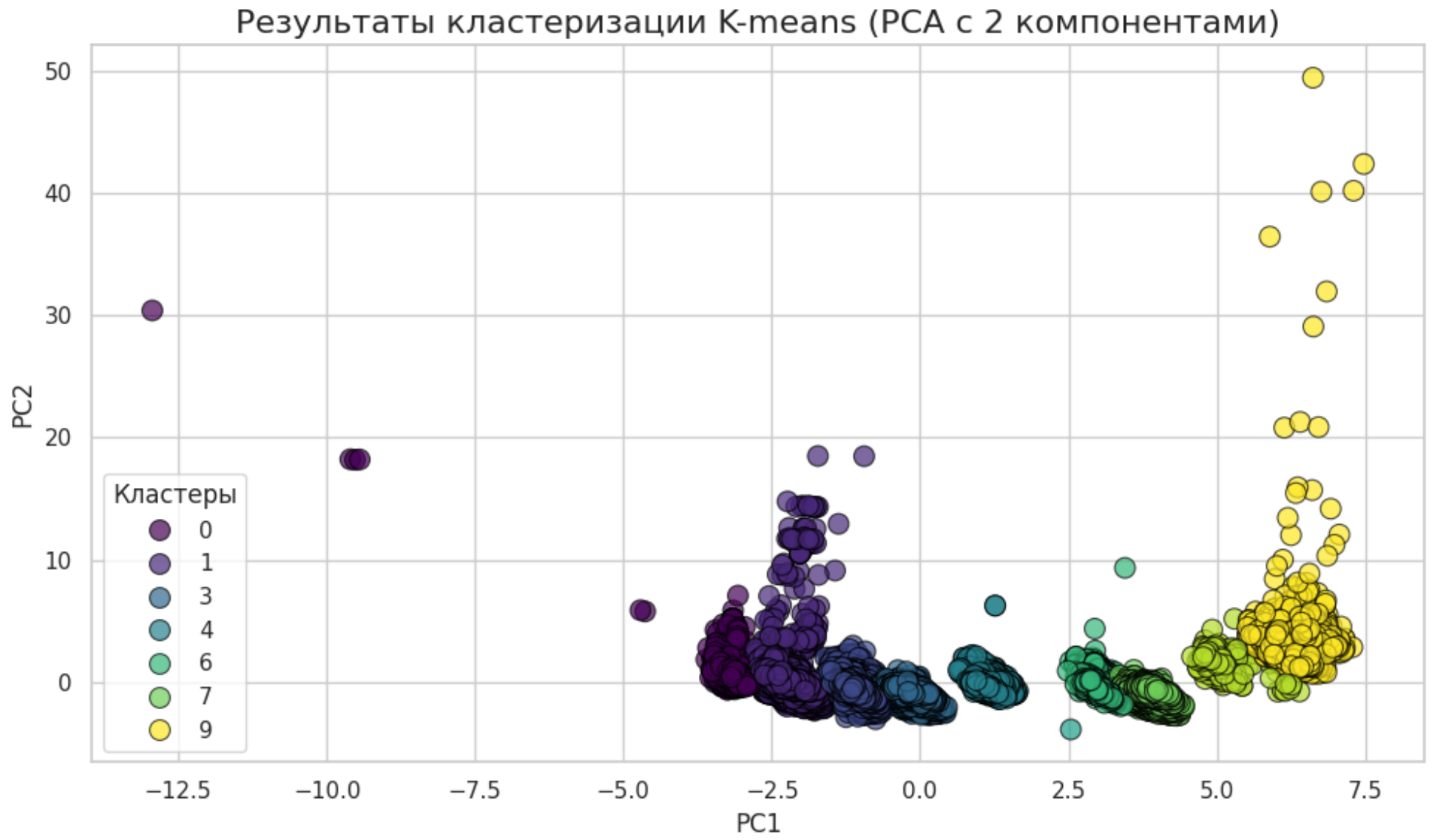
Выбор оптимального количества кластеров



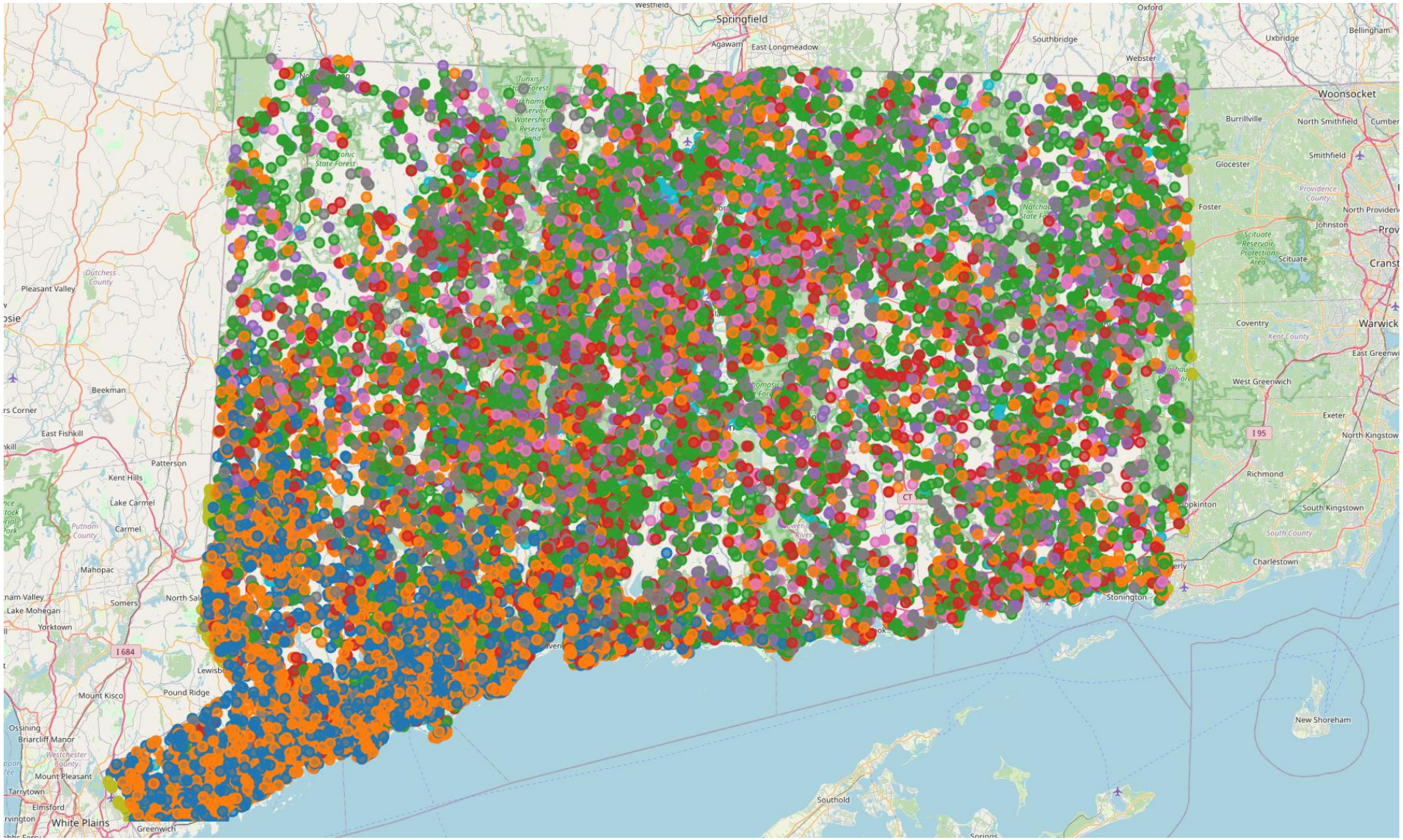
На данном графике точка локтя наблюдается при 10 кластерах. Это количество кластеров обеспечивает баланс между качеством кластеризации и сложностью модели.

Использование большего числа кластеров не приводит к значительному снижению инерции и может быть неоправданным с точки зрения интерпретируемости и вычислительных затрат.

Кластеризация K-means



Визуализация кластеризации



Визуализация кластеризации

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

Спасибо за внимание!