МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «МОСКОВСКИЙ ЭНЕРГЕТИЧЕСКИЙ ИНСТИТУТ»
ИНСТИТУТ ИНФОРМАЦИАОННЫХ И ВЫЧИСЛИТЕЛЬНЫХ ТЕХНОЛОГИЙ
КАФЕДРА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИСКУСТВЕННОГО ИНТЕЛЛЕКТА

Применение регрессионного анализа для оценки и предсказания стоимости объектов недвижимости

КУРСОВАЯ РАБОТА
по дисциплине: «Численные методы»

Выполнили:
студенты 3 курса группы А-05-22
Фролов Иван Андреевич
Сорокина Ольга Яковлевна
Преподаватель:
Амосова Ольга Алексеевна

Москва 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Прогнозирование стоимости объектов недвижимости является важной и актуальной задачей для множества областей человеческой деятельности, включая экономику, финансы, городское планирование и риэлтерскую деятельность. Современные методы обработки статистических данных, в частности, методы регрессионного анализа, позволяют с высокой точностью анализировать и предсказывать стоимость недвижимости на основе большого количества факторов. Это значительно упрощает разработку эффективных стратегий для различных участников рынка, способствует лучшему пониманию ценовых тенденций и помогает в принятии обоснованных инвестиционных решений.

Регрессионный анализ, являясь одним из ключевых инструментов в статистическом анализе данных, позволяет выявлять зависимости между целевым показателем и значимыми факторами. В случае с оценкой стоимости недвижимости такими факторами могут выступать местоположение, площадь объекта, инфраструктура, тип постройки и прочие параметры, оказывающие влияние на конечную цену. Использование регрессионных моделей предоставляет возможность не только анализировать текущие данные, но и строить прогнозы стоимости объектов в зависимости от изменяющихся условий.

Целью данного исследования является применение методов регрессионного анализа, таких как множественная линейная регрессия, деревья решений, метод случайного леса и LAD-регрессия, для построения модели прогнозирования стоимости недвижимости в Москве. Применение этих методов позволит определить наиболее значимые факторы, влияющие на ценообразование, и дать оценку точности построенных моделей.

Основные задачи исследования:

1. Провести обзор научной литературы по теме регрессионного анализа и его применения в задачах оценки недвижимости.

2. Выполнить предварительную обработку данных, включающую чистку, нормализацию и подготовку факторов.
3. Выявить и устранить выбросы в данных для повышения точности анализа.
4. Реализовать на языке Python, используя библиотеку scikit-learn, алгоритмы множественной линейной регрессии, случайного леса и LAD-регрессии применительно к поставленной задаче.
5. Провести кросс-валидацию для уточнения коэффициентов моделей.
6. Оценить адекватность моделей на тестовых данных.
7. Сравнить точность и эффективность различных моделей и методов регрессионного анализа в рамках данной задачи.

Данное исследование поможет более глубоко понять механизмы ценообразования на рынке недвижимости в Москве и определить, насколько различные факторы могут быть использованы для прогнозирования стоимости объектов.

1. Теоретические основы регрессионного анализа

Регрессия – односторонняя стохастическая зависимость, устанавливающая соответствие между случайными переменными, то есть математическое выражение, отражающее связь между зависимой переменной *y* и независимыми переменными *x* при условии, что это выражение будет иметь статистическую значимость.

$$E(y|x) = f(x) \tag{1}$$

где y − зависимая переменная, x − объясняющая переменная.

Регрессионным анализом называется поиск такой функции f, которая наилучшим образом приближает данные.

Критерием качества приближения *(целевой функцией)* обычно является среднеквадратичная ошибка:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 . \tag{2}$$

Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих.

$$y = f(x) + v, \tag{3}$$

где f - функция регрессионной зависимости, а v - аддитивная случайная величина с нулевым матожиданием.

1.1 Линейная регрессия

В **линейной регрессии** по определению объясняемая переменная зависит от другой или нескольких других независимых переменных линейно.

$$y = \omega_0 + \sum_{i=1}^{n} \omega_i x_i \tag{4}$$

Зададим модель следующим образом:

$$\mathrm{y} = X\omega + \varepsilon, \tag{5}$$

где

- $y \in \mathbb{R}^n$ – вектор с объясняемой (или целевой) переменной;
- $\omega = (1, \omega_1, \dots, \omega_n)$ – вектор параметров модели (в машинном обучении эти параметры часто называют весами);
- X – матрица наблюдений и признаков размерности n строк на m + 1 столбцов (включая фиктивную единичную колонку слева) с полным рангом по столбцам: $rank(X) = m + 1$;
- ε – случайная переменная, соответствующая случайной непрогнозируемой ошибке модели

Для отдельных наблюдений формула (5) выглядит следующим образом:

$$y_i = \sum_{j=0}^{m} \omega_i X_{ij} \; + \; \varepsilon_i$$

Также на модель накладываются следующие ограничения:

- математическое ожидание случайных ошибок равно нулю:

$$\forall \mathrm{i}\colon \mathrm{E}[\varepsilon_i] = 0$$

- дисперсия случайных ошибок одинакова и конечна:

$$Var(\varepsilon_i) = \sigma^2 < \infty$$

- случайные ошибки не коррелированы:

$$\forall i \neq j\colon \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$$

Задача регрессии заключается в поиске оценок параметров модели ω_i, которые дают минимальные отклонения между фактическими значениями зависимой переменной и восстановленными.

1.1.1. Метод наименьших квадратов

Данный метод заключается в минимизации суммы квадратов отклонений (в регрессионном анализе они чаще называются остатками регрессии), относительно параметра ω. Поскольку модель линейной регрессии задана

формулой (5), вектор оценок объясняемой переменной $\hat{y}$ и вектор остатков регрессии e будут равны:

$$\hat{y} = X\omega, e = \vec{y} - \hat{y} = \vec{y} - X\omega \quad (7)$$

Соответственно, сумма квадратов остатков регрессии будет равна:

$$RSS = e^T e = (\vec{y} - X\omega)^T(\vec{y} - X\omega). \quad (8)$$

Дифференцируя эту функцию по вектору параметров ω и приравняв производные к нулю, получим систему уравнений (в матричной форме):

$$(X^T X)\omega = X^T \vec{y}. \quad (9)$$

Отсюда найдем наш вектор ω:

$$\omega = (X^T X)^{-1} X^T \vec{y}. \quad (10)$$

При обучении весь набор данных делится на обучающую и проверочную выборки. Как следует из названия, на проверочной выборке производится проверка точности модели.

1.1.2. LAD-регрессия

Данная модель основана на методе наименьших модулей (least absolute deviation). Она используется для оценки неизвестных величин по результатам измерений, содержащих случайные ошибки, а также для приближенного представления заданной функции более простыми аппроксимациями. Похожа на линейную регрессию, но использует абсолютные величины вместо квадратов – в итоге, вместо оценивания условного математического ожидания (МНК), оценивается условная медиана. Минимизируется абсолютное расстояние между фактическим значением зависимой переменной и предсказанным, выраженное следующим функционалом:

$$d[Y, f(x)] = \sum_{i=1}^{n} |y_i - f(x_i)| \quad (11)$$

1.1.3. Разный масштаб признаков

Другой важной проблемой многомерной линейной регрессии является разнородность признаков. Если масштабы измерений признаков существенно (на несколько порядков) различаются, то появляется опасность, что будут

учитываться только «крупномасштабные» признаки. Чтобы этого избежать, делается стандартизация (нормировка) матрицы:

$$x_{ij} = \frac{x_{ij} - x_{j_{\text{ср}}}}{\sigma_j}, j = 1 \dots m, i = 1 \dots n. \quad (12)$$

где $x_{j_{\text{ср}}} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ – выборочное среднее, а $\sigma_j{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2$ – выборочная дисперсия. Причем после выполнения нормировки, ее следует применять ко всем объектам, посылаемым в качестве признака для построения предсказания.

1.1.4. Перекрестная проверка (кросс-валидация)

В случае недостаточного количества данных при делении исходной выборки на обучающую и проверочную может возникнуть проблема, когда модель плохо обучается, что влечет за собой большие ошибки прогнозирования. При использовании кросс-валидации фиксируется некоторое множество разбиений исходной выборки на две подвыборки: обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке, затем оценивается его средняя ошибка на объектах контрольной подвыборки. Оценкой кросс-валидации называется средняя по всем разбиениям величина ошибки на контрольных подвыборках. Существует несколько способов реализации кросс-валидации, отличающиеся способом разбиения выборки, рассмотрим способ, использованный в исследовании. Изначальная выборка X разбивается на k примерно одинаковых по длине частей, таким образом, что: $X = X_1 \cup X_2 \cup \dots \cup X_k$. После этого модель начинает обучаться на $k-1$ блоке, а не вошедший в обучающую выборку блок, становится проверочным, и так k раз. Обычно принято делить выборку на 10 частей. Если выборка независима, то средняя ошибка кросс-валидации даёт несмещённую оценку вероятности ошибки. Это выгодно отличает её от средней ошибки на обучающей выборке, которая может оказаться смещённой (оптимистически заниженной) оценкой вероятности ошибки, что связано с явлением переобучения.

1.2. Метод случайного леса

1.2.1. Решающие деревья

Решающее дерево (Decision tree) – метод решения задачи обучения с учителем, основанный на том, как решает задачи прогнозирования человек. В общем случае — это k-ичное дерево с решающими правилами в нелистовых вершинах (узлах) и некоторое заключение о целевой функции в листовых вершинах (прогнозом).

Схеме принятия решений соответствует связный ориентированный ациклический граф – ориентированное дерево. Дерево включает в себя корневую вершину, инцидентную только выходящим рёбрами, внутренние вершины (узлы), инцидентную одному входящему ребру и нескольким выходящим, и листья – концевые вершины, инцидентные только одному входящему ребру.

Каждый узел и корень содержат решающее правило.

Решающее правило — некоторая функция от объекта, позволяющее определить, в какую из дочерних вершин нужно поместить рассматриваемый объект. В листовых вершинах могут находиться разные объекты: класс, который нужно присвоить попавшему туда объекту (в задаче классификации), вероятности классов (в задаче классификации), непосредственно значение целевой функции (задача регрессии). Как вариант можно использовать следующее решающее правило: $\beta(x, j, t) = [x_j < t]$, где t некоторая константа.

Решающие правила разбивают с помощью рекурсивного бинарного разбиения, пространство на некоторое количество непересекающихся подмножеств $\{J_1, J_2, J_3, \dots, J_n\}$, и в каждом подмножестве J_j выдают константный прогноз ω_j. Значит, соответствующий алгоритм можно записать аналитически:

$$f(x) = \sum_{j=1}^{n} \omega_i I(x \in \mathrm{J_j}) \qquad (13)$$

Алгоритм построения дерева:

1. Проверить критерий останова алгоритма. Если он выполняется, выбрать для узла выдаваемый прогноз, что можно сделать несколькими способами (Критерий останова может быть разнообразный: ограничение

максимальной глубины дерева, ограничение минимального числа объектов в листе, ограничение максимального количества листьев в дереве и т.д.).

2. Иначе требуется разбить множество на несколько не пересекающихся. В общем случае в вершине t задаётся решающее правило $Q_t(x)$, принимающее некоторый диапазон значений. Этот диапазон разбивается на R_t непересекающихся множеств объектов, $S_1, S_2, \dots, S_{R_t}$, где R_t – количество потомков у вершины, а каждое S_i – это множество объектов, попавших в i-го потомка.

3. Множество в узле разбивается согласно выбранному правилу, для каждого узла алгоритм запускается рекурсивно.

1.2.2 Алгоритм случайного леса

В исследовании будет использоваться алгоритм, являющийся надстройкой над решающими деревьями, а именно алгоритм случайного леса, один из лучших для прогнозирования, но имеющий следующий недостаток: увидеть явный вид обученной модели невозможно.

Основная идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество, однако при использовании большого количества деревьев качество повышается. Существует много версий алгоритма, но выделяют две основных, это: для количественных переменных y_i, и для качественных переменных y_i. Рассмотрим алгоритм построения леса:

1. Выбирается подвыборка из обучающей выборки размера l (с возможностью повторного выбора элементов) – по ней строится дерево (для каждого дерева – своя подвыборка);

2. Для построения каждого расщепления в дереве просматриваем k случайных признаков из всех (для каждого нового расщепления — свои случайные признаки);

3. Выбираем наилучший признак и расщепление по нему (по заранее заданному критерию, например по минимизированию RSS). Дерево строится,

как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса);

4. Вычисляем итоговое предсказание $\alpha(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x)$, где $b_i(x)$ – предсказание в i – ом дереве.

Рекомендуется в задачах регрессии брать количество случайных признаков $k = \frac{n}{3}$, где n – количество всех признаков. Этот выбор основывается на эмпирических наблюдениях и предназначен для достижения баланса между смещением и дисперсией в модели случайного леса.

1.3. Удаление выбросов из набора данных

1.3.1. Точки с избыточным влиянием на модель

Говоря о необычных наблюдениях в контексте регрессионного анализа, можно выделить следующие три ситуации:

- наблюдение представлено необычным сочетанием значений предикторов. Это наблюдения, где комбинация значений независимых переменных (предикторов) выделяется на фоне других данных. Например, в выборке может быть точка с крайне высоким значением одного предиктора и низким значением другого, что не характерно для остальных наблюдений.
- наблюдение не согласуется с рассматриваемой моделью, т.е. является выбросом;
- наблюдение оказывает существенное влияние на оценки параметров модели; другими словами, удаление такого влиятельного наблюдения из выборки приведет к значительному изменению предсказываемых моделью значений.

Необычные наблюдения могут оказывать существенное влияние на качество модели (как с точки зрения статистической значимости ее параметров, так и с точки зрения ее предсказательной силы), в связи с чем выявление таких наблюдений является важной частью диагностики регрессионных моделей.

Имеется возможность выразить потенциал воздействия количественно. Распространенным подходом является расчет матрицы влияния (hat matrix):

$$\mathrm{H} = X(X^T X)^{-1} X^T \tag{14}$$

Диагональные элементы этой матрицы h_{ii} называются значениями хэта (*hat values*). Они измеряют, насколько далеко каждая точка находится от центра распределения предикторов, а также потенциал влияния этой точки на модель.

Оказывается, что сумма диагональных элементов матрицы проекции равна числу коэффициентов регрессионного уравнения, включая свободный член. Соответственно, среднее значение h_{ii} можно рассчитать как p/n, где p – число коэффициентов регрессионного уравнения, n – количество наблюдений. Отсюда вытекает эмпирическое правило, позволяющее судить о том, оказывает ли некоторое наблюдение существенное влияние на параметры модели - значения $h_{ii} > 2p/n$ являются достаточно большими, чтобы считать соответствующие наблюдения стоящими внимания.

1.3.2. Выбросы

Выброс – это наблюдение с большим остатком, возникающим из-за того, что соответствующее выборочное значение зависимой переменной y_i значительно отличается от предсказанного значения.

На практике работать с изначальными значениями выбросов оказывается проблематично, поэтому прибегают к различным стандартизациям. Часто для диагностики линейных моделей, чьи параметры оцениваются по методу наименьших квадратов, используются следующие два типа остатков:

Studentized Residuals (Стандартизированные остатки):

$$r_i = \frac{\varepsilon_i}{S_\varepsilon \sqrt{1 - h_{ii}}} \tag{15}$$

где ε_i – остаток i – го наблюдения, S_ε - стандартное отклонение всех остатков модели, а h_{ii} – показатель потенциала воздействия i – го наблюдения на коэффициенты модели (см. 1.3.1. формула 14).

Одним из важных недостатков стандартизованных остатков является тот факт, что любое значение r_i и S_ε не является независимым, затрудняя

формальную проверку статистической гипотезы о том, что некоторое i-е наблюдение не является выбросом.

Для устранения указанного недостатка используют Studentized residuals - стьюдентизированные остатки:

$$t_i = \frac{\varepsilon_i}{S_{\varepsilon_{(-t)}}\sqrt{1 - h_{ii}}} \tag{16}$$

где $S_{\varepsilon_{(-t)}}$ - стандартное отклонение, которое рассчитывается по остаткам модели, подогнанной после исключения из данных i – го наблюдения.

Стьюдентизированные остатки имеют распределение Стьюдента с $n - p - 1$ степенями свободы. Соответственно, мы можем использовать квантили этого распределения для проверки того, насколько статистически значимо определенное наблюдение является выбросом. Если вычисленное значение $t_i \geq t_{\text{кр}}$, где $t_{\text{кр}}$ − это квантиль распределения Стьюдента с $n - p - 1$ степенью свободы и уровнем значимости $\alpha = 0.05$, то данное наблюдение рассматривается как выброс и подлежит удалению из выборки.

1.4. Корреляционная матрица и анализ признаков

Анализ признаков играет ключевую роль в построении эффективной регрессионной модели. Корреляционная матрица является одним из инструментов, который используется для выявления линейных взаимосвязей между числовыми признаками. Она представляет собой квадратную таблицу, где на пересечении строк и столбцов расположены коэффициенты корреляции между соответствующими переменными.

1.4.1. Коэффициент корреляции Пирсона

Основным методом вычисления корреляции является использование коэффициента Пирсона. Этот коэффициент измеряет силу и направление линейной связи между двумя переменными и выражается следующим образом:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{17}$$

где

- x_i, y_i – значения двух переменных;

- $\bar{x}, \bar{y}$ – их средние значения.

Коэффициент принимает значения от -1 до 1:

- $r = 1$: Полная положительная линейная зависимость.
- $r = -1$: Полная отрицательная линейная зависимость.
- $r = 0$: Отсутствие линейной зависимости.

1.4.2. Применение корреляционной матрицы

Корреляционная матрица используется для следующих целей:

1. **Определение мультиколлинеарности:** Если два признака имеют высокую корреляцию ($|r| > 0.8$), то они могут быть избыточными, что приведет к мультиколлинеарности. В таких случаях рекомендуется:
    - Исключить один из сильно коррелированных признаков.
    - Провести их объединение (например, методом главных компонент).
2. **Выявление слабокоррелированных признаков:** Признаки с низкой корреляцией с целевой переменной могут оказаться менее информативными и быть исключены из модели для снижения её сложности.
3. **Интерпретация зависимости:** Анализ корреляции позволяет выявить признаки, которые оказывают наибольшее влияние на целевую переменную. Это помогает лучше понять структуру данных.

1.5. Оценка качества модели регрессии

1.5.1. Коэффициент детерминации

Коэффициент детерминации показывает, какая доля изменения исследуемого признака учтена в модели. Коэффициент детерминации R^2 может принимать значения от 0 до 1. Чем ближе коэффициент детерминации R^2 к единице, тем лучше качество модели.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}\left(y_i - y_{\text{ср}}\right)^2} \quad (20)$$

где

- $RSS = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$ – сумма квадратов регрессионных остатков;
- $TSS = \sum_{i=1}^{n}\left(y_i - y_{\text{ср}}\right)^2$ – общая дисперсия;
- $y_i, \widehat{y_i}$ – соответственно, фактические и расчетные значения объясняемой переменной;
- $y_{\text{ср}} = \frac{1}{n}\sum_{i=1}^{n} y_i$ – выборочное среднее.

1.5.2. Функции ошибок

Среднеквадратическая ошибка

Точность подгонки модели под данные оценивается с помощью среднеквадратической ошибки по формуле (4). Используется для выбора среди нескольких моделей, с последующим выбором модели с наименьшей данной ошибкой. Соответственно, алгоритм, рассчитывая регрессионную модель, стремится минимизировать этот коэффициент.

Средний модуль отклонения

Также были рассмотрены средние модули отклонения предсказанной величины от настоящего значения.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

В данной величине отображается усредненное значение отклонений фактических данных от предсказанных. Недостаток данной ошибки состоит в непонимании насколько велико значение относительно имеющихся данных.

1.5.3. Grid Search для оптимизации гиперпараметров

Одним из важных шагов в процессе построения и улучшения моделей машинного обучения является настройка гиперпараметров. Grid Search (поиск по сетке) является одним из самых популярных методов для систематического подбора оптимальных значений гиперпараметров модели. Он заключается в том, что на предварительном этапе определяется сетка возможных значений гиперпараметров, и затем для каждой комбинации параметров модель обучается и оценивается с использованием кросс-валидации.

Этот метод позволяет не только найти наиболее подходящие гиперпараметры, но и избежать проблемы переобучения, так как оценка производительности модели производится на различных подмножествах данных. Применяя Grid Search, можно настроить такие параметры, как количество деревьев в методе случайного леса, максимальную глубину деревьев, минимальное количество образцов для разделения узла и другие ключевые параметры.

Процесс работы с Grid Search включает несколько шагов:

1. **Определение сетки гиперпараметров**: На этом этапе выбираются параметры модели, которые будут изменяться, и для каждого параметра определяются возможные значения.

2. **Обучение модели для каждой комбинации параметров**: Для каждой комбинации гиперпараметров модель обучается и проходит через процесс кросс-валидации.

3. **Оценка результатов**: После завершения обучения модели для каждой комбинации гиперпараметров производится оценка её производительности с использованием выбранной метрики, например, коэффициента детерминации R^2 или среднеквадратичной ошибки MSE.

4. **Выбор наилучшей модели**: Комбинация гиперпараметров, которая дает наилучший результат, считается оптимальной и используется для финального обучения модели.

Для поиска наилучших значений этих гиперпараметров с использованием Grid Search можно определить сетку параметров и обучить модель с кросс-валидацией. Это позволяет найти оптимальные параметры, которые обеспечат наилучшую производительность модели.

2. Решение задачи прогнозирования стоимости недвижимости в Москве в зависимости от различных факторов

2.1. Описание набора и задачи исследования

Анализируемый набор данных содержит информацию о продаже недвижимости в Москве в 2014 году. Количество данных в наборе – 8894. Он включает как характеристики объектов недвижимости, так и данные об окружающих районах. Описание признаков набора данных представлено ниже.

Информация об объекте недвижимости:

- **full_sq** — общая площадь в м², включая лоджии, балконы и другие дополнительные площади;
- **life_sq** — жилая площадь в м², без учета лоджий и балконов;
- **floor** — этаж, на котором расположена квартира;
- **max_floor** — общее количество этажей в здании;
- **material** — материал здания, представленный следующими категориями:
    - панель (panel),
    - кирпич (brick),
    - дерево (wood),
    - монолит (mass concrete),
    - блоки (breezeblock),
    - кирпично-монолитное (mass concrete plus brick);
- **build_year** — год постройки дома;
- **num_room** — количество комнат в квартире;
- **metro_min_avto** — время в минутах до ближайшей станции метро на автомобиле;
- **metro_km_avto** — расстояние в километрах до ближайшей станции метро на автомобиле;

- **metro_min_walk** — время в минутах до ближайшей станции метро пешком;
- **metro_km_walk** — расстояние в километрах до ближайшей станции метро пешком;
- **mkad_km** — расстояние в километрах до МКАД (Московская кольцевая автодорога);
- **kremlin_km** — расстояние в километрах до Кремля;
- **green_part_1000** — процент зеленых зон в радиусе 1 км;
- **prom_part_1000** — процент промышленных зон в радиусе 1 км;
- **office_count_1000** — количество офисных зданий в радиусе 1 км;
- **trc_count_1000** — количество торгово-развлекательных центров в радиусе 1 км;
- **leisure_count_1000** — количество мест отдыха в радиусе 1 км;
- **price_doc** — цена продажи объекта недвижимости (целевой признак).

Информация о районе:

- **sub_area** — название района;
- **area_m** — площадь района в м²;
- **green_zone_part** — доля зеленых зон в районе;
- **industri_part** — доля промышленных зон в районе;
- **preschool** — количество детских садов в районе;
- **school** — количество школ в районе;
- **healthcare** — количество медицинских центров в районе;
- **radiation** — наличие радиоактивных отходов в районе (1 — есть, 0 — нет);
- **detention** — наличие тюрем в районе (1 — есть, 0 — нет);
- **young** — количество людей, не достигших трудоспособного возраста;
- **work** — количество людей трудоспособного возраста;
- **elder** — количество людей пенсионного возраста;

- **0_6_age** — количество людей в возрасте до 6 лет;
- **7_14_age** — количество людей в возрасте от 7 до 14 лет.

Целевым признаком является **price_doc**, который представляет собой цену продажи объекта недвижимости в рублях.

Целью данного исследования является разработка модели, которая бы точно предсказывала стоимость квартир в Москве.

Анализ и моделирование данных о недвижимости имеют большое практическое значение, так как позволяют как покупателям, так и продавцам принимать более обоснованные решения. Особенно это может быть полезно при выборе факторов, на основе которых следует формировать цены на недвижимость.

2.2 Первичная обработка данных

2.2.1. Обработка NaN значений.

В процессе предварительной обработки данных был проведен анализ пропущенных значений (NaN). В наборе данных обнаружены следующие пропущенные значения:

- В столбце build_year отсутствуют данные для 35 объектов недвижимости;
- В столбцах metro_min_walk и metro_km_walk пропущены значения для 11 объектов недвижимости.

	Count NaN values
build_year	35
metro_min_walk	11
metro_km_walk	11

Пропущенные значения в столбце build_year(год постройки)

При анализе статистических характеристик столбца build_year были обнаружены потенциальные выбросы и неправильные значения. Рассмотрим основные статистические показатели:

- Минимальное значение: 0;
- Максимальное значение: 4965.
- Ненулевое число пропущенных значений

Такие значения явно не соответствуют реальным данным, поскольку год постройки здания не может быть равен 0 или 4965. Вероятнее всего, это результат ошибок при сборе данных или некорректных записей.

Так как эти ошибки сильно влияют как на восприятие, так и на перспективы нашей работы, проведем процесс обработки.

Просто удалить эти данные из таблицы мы не можем, так как набор станет слишком маленьким.

Чтобы решить эту проблему:

- Мы определим минимальный год постройки, основываясь на материале строительства;
- Значения ниже и выше этих ограничений будут заменены на среднее значение года постройки по материалу;
- Использование квантилей в данном случае будет не совсем корректным — мы потеряем достоверные данные, находящиеся ниже квантиля.

Замена будет происходит исходя их фактов:

- Первый панельный дом в СССР был построен после войны в 1948 году в Москве.
- Впервые крупные шлакоблоки в России начали использовать в 1927 году.

- Использование цемента началось очень давно, поэтому здания, имеющие меньшие значения, чем нормальные минимальные значения, мы заменим на средние значения, основанные на имеющихся данных,
    - Нижнюю границу выберем вручную из отсортированной по году выборки.
- Использование кирпича также началось слишком давно, чтобы здания могли иметь меньшие значения, чем нормальные минимальные значения, поэтому мы заменим их средними значениями на основе имеющихся данных,
    - Нижнюю границу выберем вручную из отсортированной по году выборки.

Обработка пропущенных значений, относящихся к признакам, связанным с метро

Заменим NaN на медианные значения в данных по текущему признаку.

В итоге мы избавились от всех NaN значений.

2.3. Визуализация

2.3.1. Одномерный анализ данных

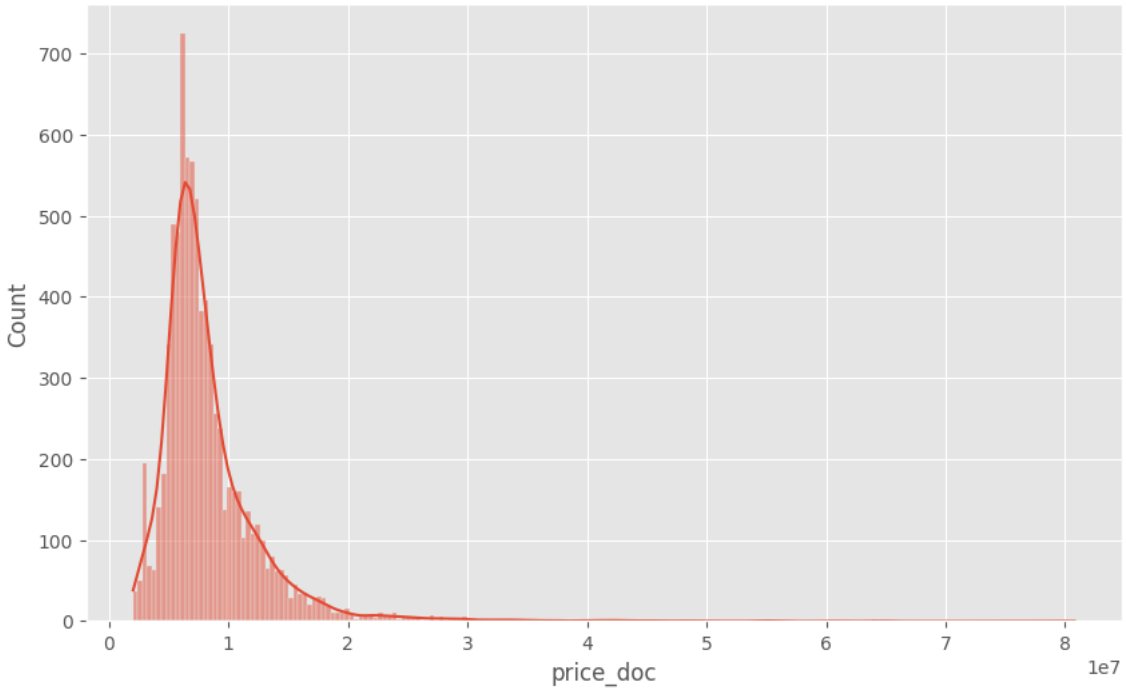
Построим гистограмму, чтобы посмотреть на распределение Y

Значение коэффициента асимметрии распределения значений в столбце price_doc DataFrame numeric_cols (набор данных, где фигурируют только количественные признаки) равно 3.509339.

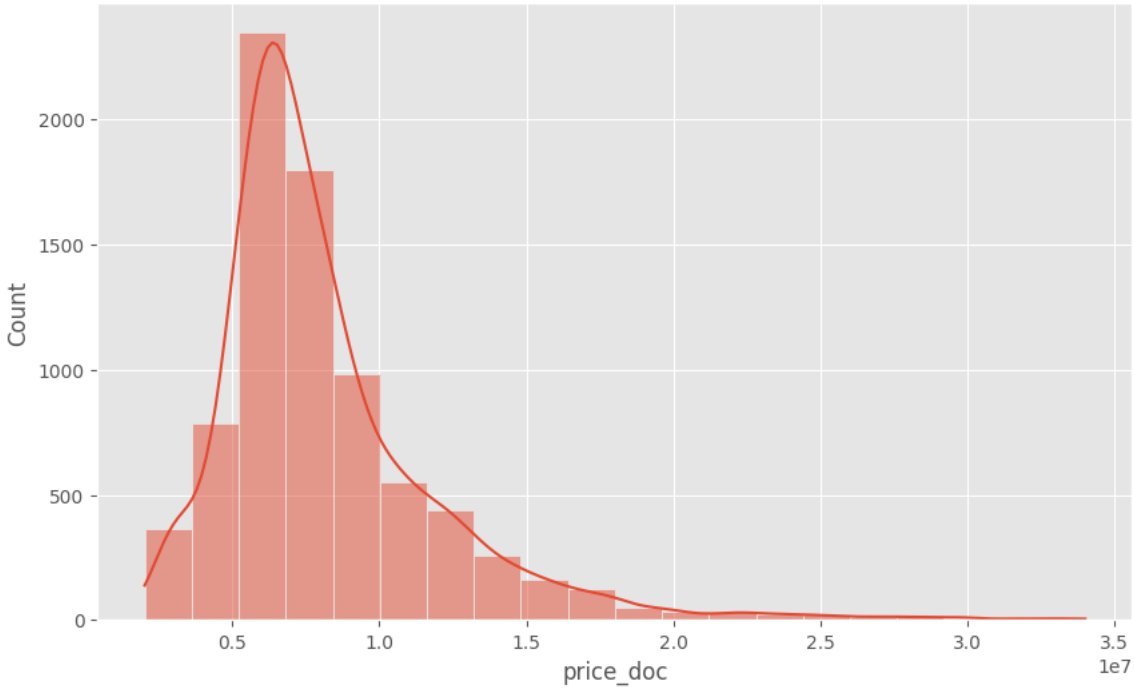
По графику и коэффициенту асимметрии видно, что распределение имеет правостороннюю асимметрию.

Определив значения выбросов, получим, что в наборе много квартир стоимостью ровно 2 млн.руб. Это выглядит странно, поэтому удалим эти данные из набора.

Получили новое распределение:



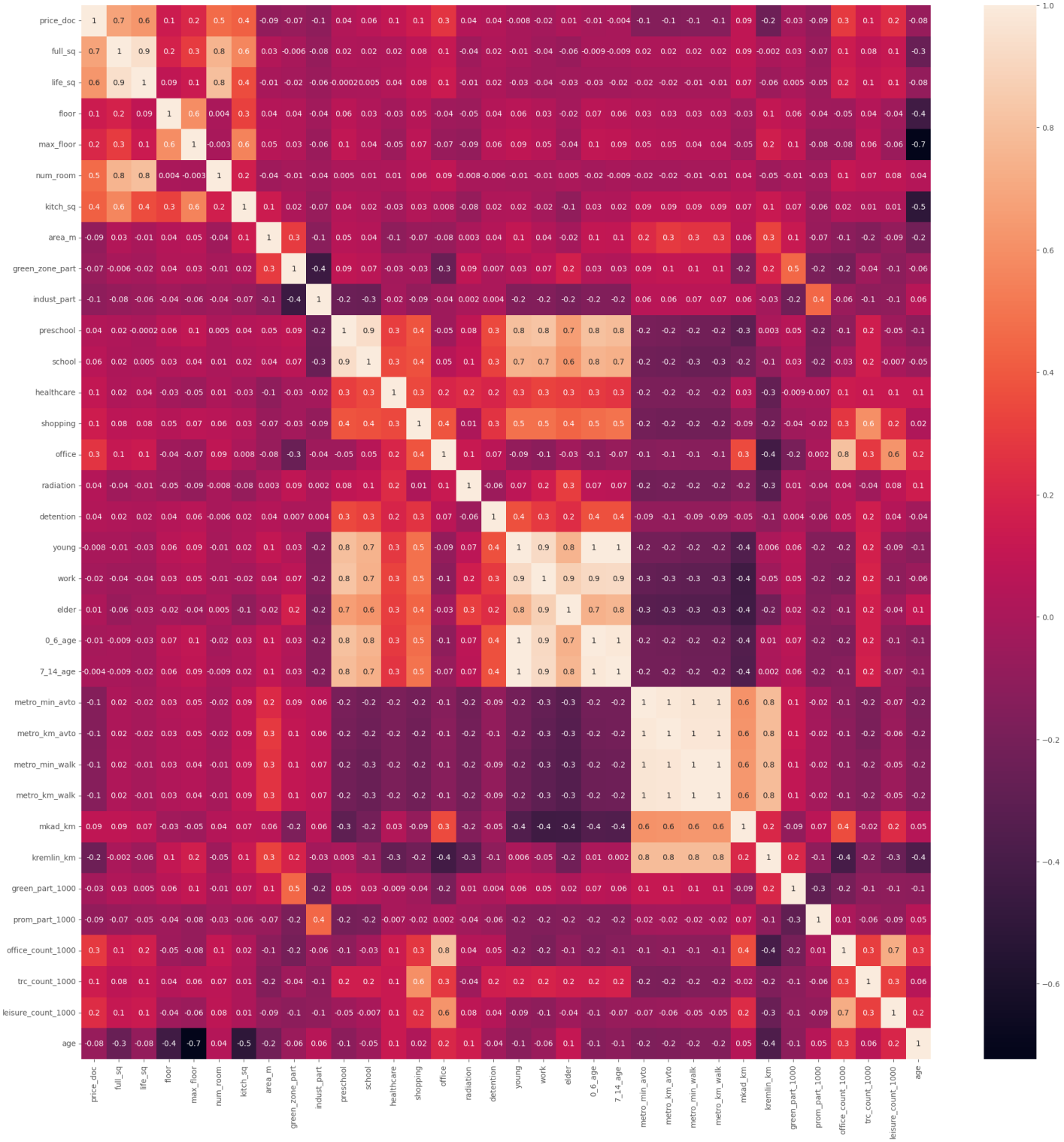
Выполним фильтрацию выбросов в данных, используя метод межквартильного размаха (IQR — Interquartile Range)



Получили более или менее нормальное распределение.

2.3.1. Многомерный анализ данных

Для оценки взаимосвязи между числовыми признаками была построена тепловая карта корреляционной матрицы. На графике отображены коэффициенты корреляции между всеми числовыми переменными в наборе данных. Использование аннотирования значений на тепловой карте позволяет наглядно видеть степень и направление корреляции между признаками.

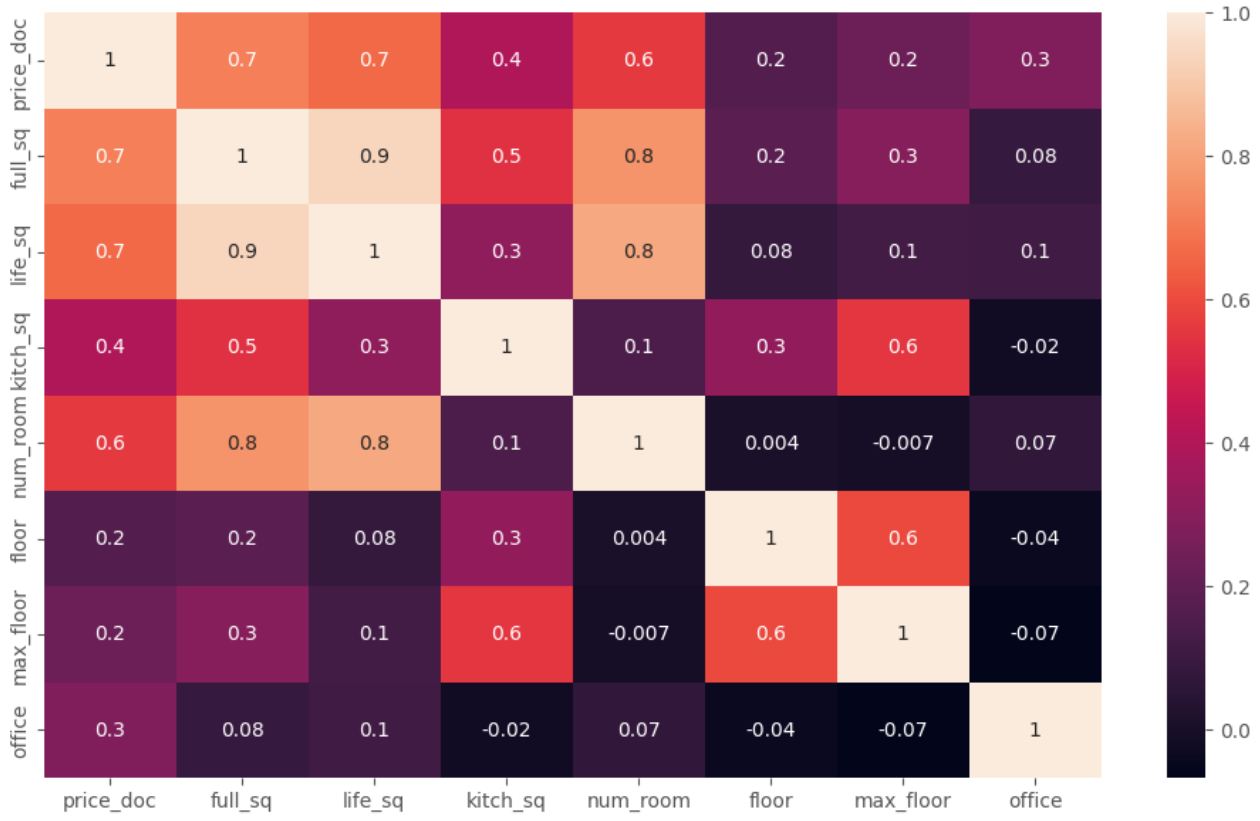


Промежуточные выводы

Наибольшая корреляция с результирующей характеристикой price_doc показана признаками:

- **full_sq** - площадь квартиры и соотносящиеся с ней life_sq и kitchen_sq
- **office** - количество офисов в этом районе.

Выполним анализ корреляции между столбцами данных price_doc, full_sq, life_sq, kitch_sq, num_room, floor, max_floor, office, отобразим тепловую карту и выведем коэффициент корреляции Пирсона для каждого столбца с price_doc.



Результаты вычислений коэффициента корреляции Пирсона для различных признаков с целевой переменной **price_doc** и другими числовыми признаками можно представить следующим образом:

- **price_doc**: коэффициент корреляции 1.0, p-значение 0.0 (полная положительная корреляция с собой).

- **full_sq**: коэффициент корреляции 0.715, p-значение 0.0 (сильная положительная корреляция с **price_doc**).
- **life_sq**: коэффициент корреляции 0.663, p-значение 0.0 (средняя положительная корреляция с **price_doc**).
- **kitch_sq**: коэффициент корреляции 0.396, p-значение 7.74e-299 (слабая положительная корреляция с **price_doc**).
- **num_room**: коэффициент корреляции 0.559, p-значение 0.0 (умеренная положительная корреляция с **price_doc**).
- **floor**: коэффициент корреляции 0.165, p-значение 6.18e-50 (очень слабая положительная корреляция с **price_doc**).
- **max_floor**: коэффициент корреляции 0.229, p-значение 1.72e-95 (слабая положительная корреляция с **price_doc**).
- **office**: коэффициент корреляции 0.275, p-значение 4.14e-139 (слабая положительная корреляция с **price_doc**).

P-значение (p-value) указывает на статистическую значимость корреляции. В данном случае, для всех признаков, за исключением **price_doc**, p-значения равны нулю или очень близки к нулю, что означает, что все найденные корреляции статистически значимы.

Как видно, все выбранные параметры имеют корреляцию с целевой переменной Y, но стоит обратить внимание на переменные num_room, floor/max_floor, life_sq, kitch_sq.

Кроме того, следует обратить внимание на сильную взаимосвязь между этими параметрами.

В данном случае потребуется выбрать дополнительные параметры, которые могут быть использованы для построения модели.

2.4. Выбор дополнительных параметров

Учитывая, что другие параметры в наборе данных имеют взаимную корреляцию, существует основание для создания коэффициента с весами для повышения точности модели.

Лучшие идеи:

- Данные о привлекательности района.
- Средняя стоимость за квадратный метр в районе / Средняя стоимость по району. Поскольку использование только стоимости квадратного метра в квартире будет некорректным, при наличии таких данных модель будет излишней.
- Данные о транспортной системе.
- Оценка привлекательности в радиусе 1 км от квартиры.
- Коэффициент качества дома.
- Коэффициент достаточности инфраструктуры.

2.4.1. Данные о привлекательности района

Признаки, которые так или иначе говорят о привлекательности района: area_m, green_zone_part, indust_part, preschool, school, healthcare, radiation, detention, young, work, elder, 0_6_age, 7_14_age, shopping, office.

Построив гистограммы для каждого признака (см. в приложении), можно сделать вывод, что не все данные распределены нормально. И при их нормализации значительная часть выборки будет потеряна, и обычные показатели для большинства этих факторов будут утрачены.

Некоторые данные обрежем с использованием квантилей:

- area_m
- office
- green_zone_part

- indust_part

Результаты корреляции Пирсона для признаков представлены ниже:

- **area_m**: коэффициент корреляции = -0.0964, p-значение = 1.6368e-17
- **green_zone_part**: коэффициент корреляции = -0.0620, p-значение = 4.4825e-08
- **indust_part**: коэффициент корреляции = -0.0905, p-значение = 1.3212e-15
- **preschool**: коэффициент корреляции = 0.0283, p-значение = 0.0127
- **school**: коэффициент корреляции = 0.0536, p-значение = 2.2740e-06
- **healthcare**: коэффициент корреляции = 0.0715, p-значение = 2.8732e-10
- **radiation**: коэффициент корреляции = 0.0187, p-значение = 0.0996
- **detention**: коэффициент корреляции = 0.0290, p-значение = 0.0106
- **young**: коэффициент корреляции = -0.0136, p-значение = 0.2314
- **work**: коэффициент корреляции = -0.0234, p-значение = 0.0393
- **elder**: коэффициент корреляции = 0.0067, p-значение = 0.5568
- **0_6_age**: коэффициент корреляции = -0.0118, p-значение = 0.2977
- **7_14_age**: коэффициент корреляции = -0.0132, p-значение = 0.2429
- **shopping**: коэффициент корреляции = 0.1218, p-значение = 4.5165e-27
- **office**: коэффициент корреляции = 0.2283, p-значение = 1.9079e-92

Как видно из p-значения, только несколько параметров действительно не имеют корреляции с целевой переменной.

Исключаем ненужные данные.

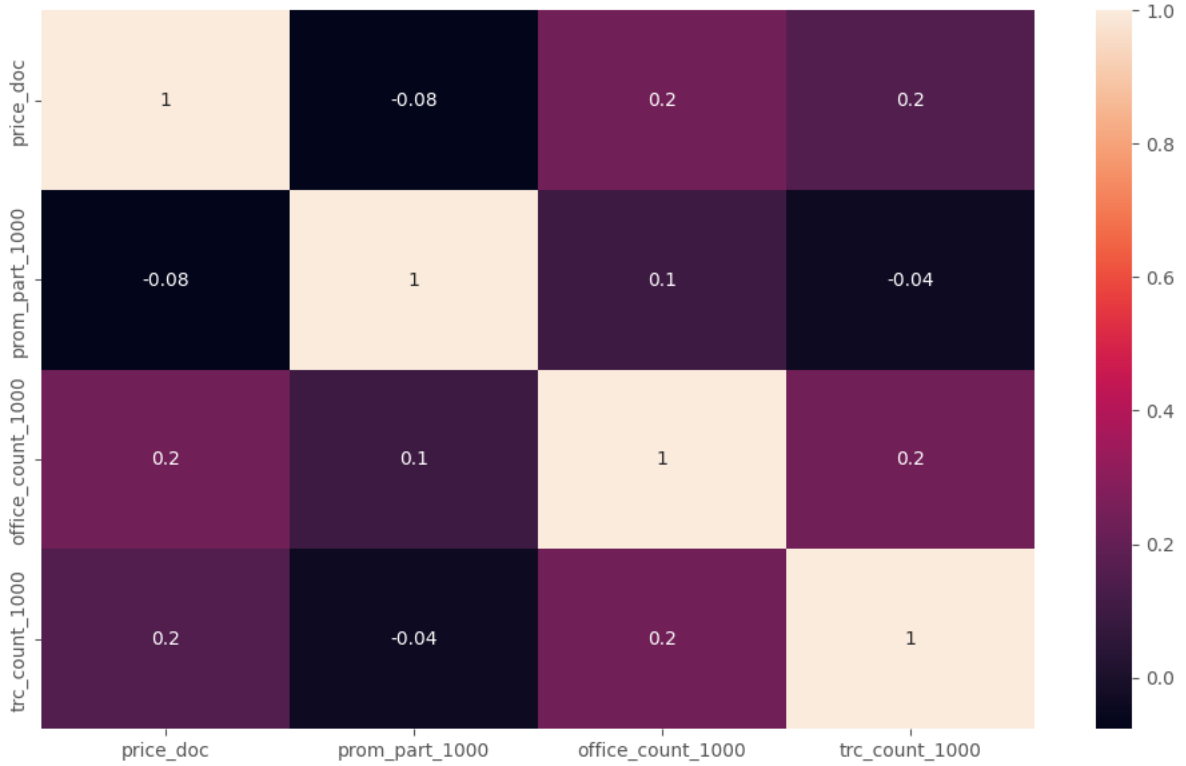
Также исключаем параметры "preschool", "young", "elder", "children" и "detention", так как они показывают сильную взаимосвязь между собой.

В итоге делаем вывод, что все остальные признаки привлекательности района важны.

2.4.2. Коэффициент привлекательности в радиусе километра от квартиры.

В наборе есть признаки green_part_1000, prom_part_1000, office_count_1000, trc_count_1000, leisure_count_1000, которые говорят о привлекательности района в радиусе одного километра.

Построив гистограммы (см. приложение), сделаем вывод, что исключение нулевых и противоположных минимальных значений приведет к сокращению выборки и не покажет взаимосвязь между ценой и километровыми радиусами, где нет торговых центров, парков и т.д.



Нам придется удалить все признаки, так как они показывают слишком слабую корреляцию и также коррелируют между собой.

Вместо этого мы создадим новый признак — коэффициент привлекательности в радиусе километра от квартиры.

Для этого обучим линейную регрессию на признаках prom_part_1000, trc_count_1000, office_count_1000 и найдем веса и свободный член.

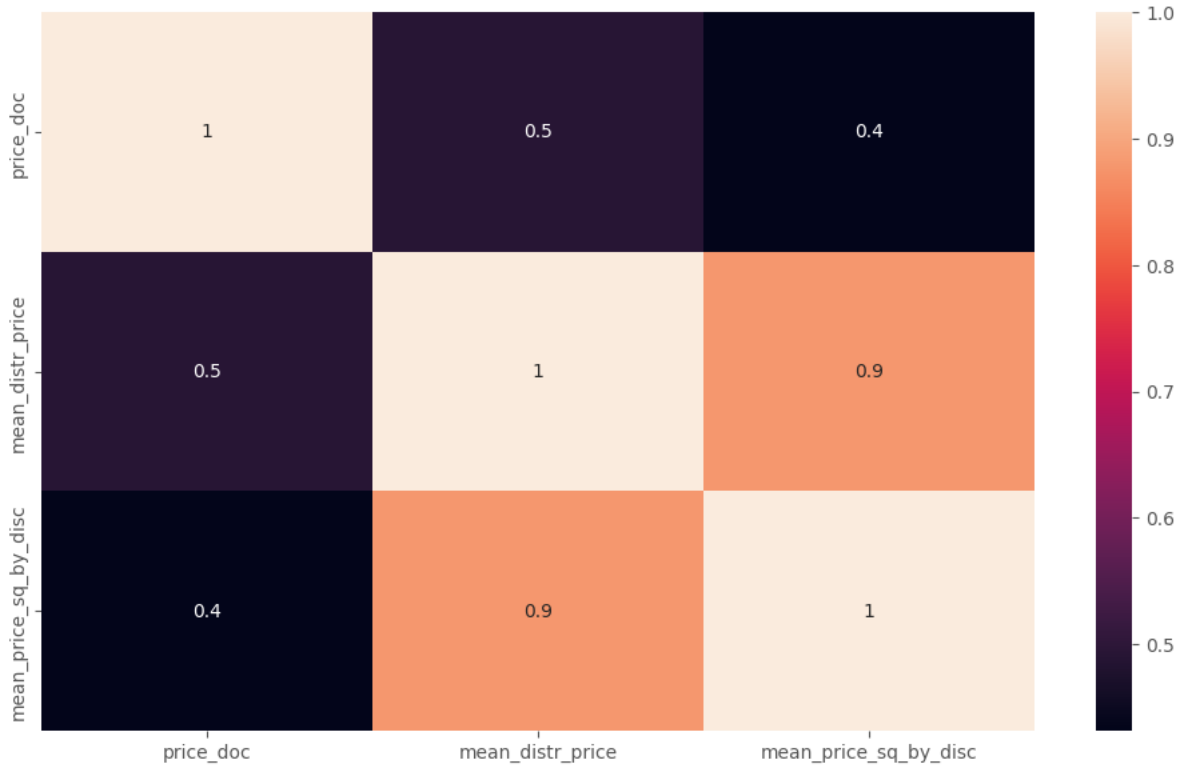
Для создания нового признака km_score было выполнено вычисление линейной комбинации независимых переменных, умноженных на их соответствующие коэффициенты с добавлением свободного члена.

Результат корреляционного анализа Пирсона для пары признаков price_doc и km_score показывает, что коэффициент корреляции составляет 0.244, что свидетельствует о слабой положительной корреляции между этими признаками. P-значение при этом равно 8.28e-106, что значительно меньше стандартного порога значимости (обычно 0.05), указывая на статистическую значимость этой корреляции.

Это означает, что корреляция между переменными статистически значима, хотя и слабая. В зависимости от контекста исследования и цели модели, такую переменную можно оставить, особенно учитывая тот факт, что она имеет практическую значимость.

2.4.3. Средняя цена квартир на районе.

Мы рассчитали среднюю стоимость квартиры и среднюю площадь, а затем добавили их в качестве признаков. Из таблицы корреляции нетрудно заметить, что оба показателя являются значимыми и имеют довольно сильную корреляцию с результирующим признаком.



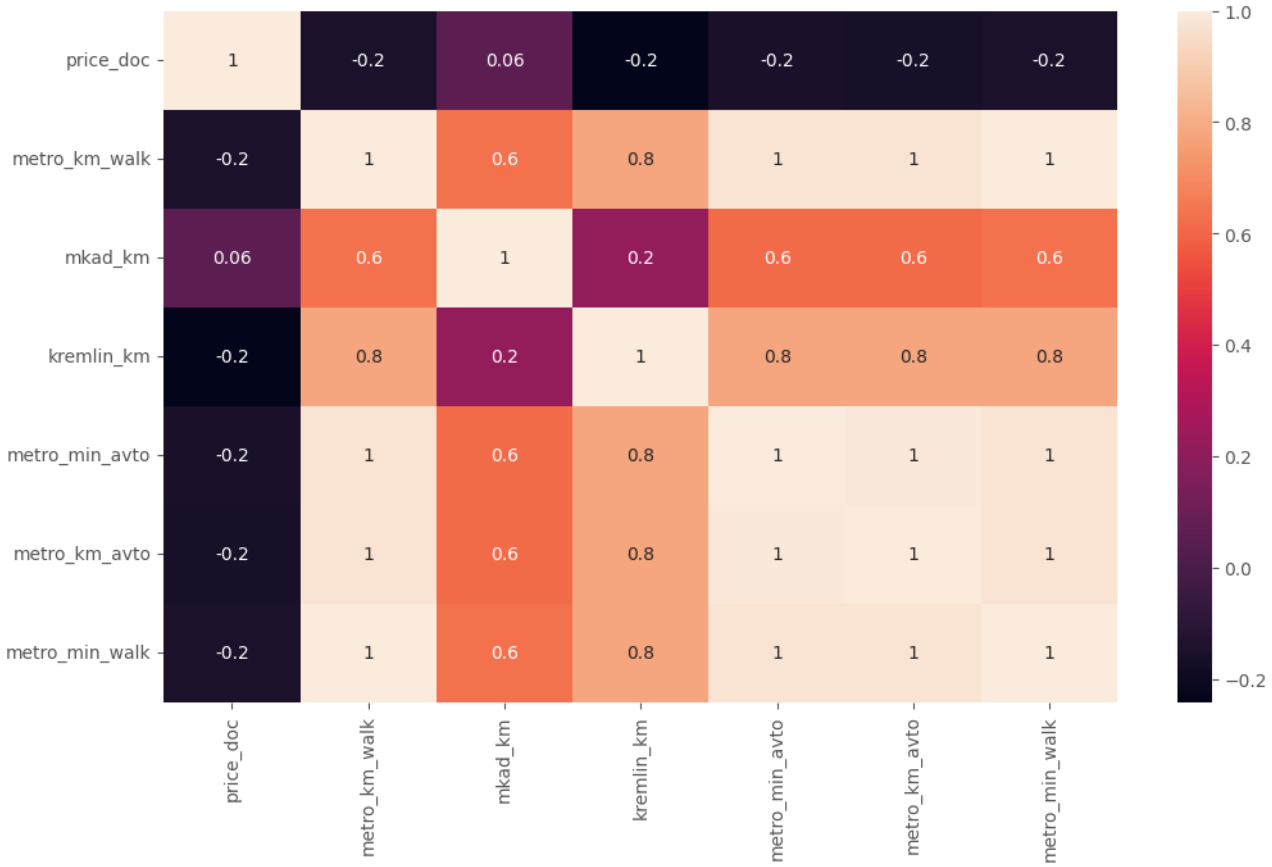
Однако эти признаки имеют сильную корреляцию друг с другом, поэтому убираем среднюю цену за квадратный метр, так как корреляция средней цены квартиры в текущем районе выше.

Несмотря на вышеуказанное решение, средняя цена за квадратный метр остаётся более полезным признаком, особенно в долгосрочной перспективе. Поэтому можно пожертвовать немного качеством модели сейчас, оставив этот признак, так как он будет играть ключевую роль в будущем. В итоге оставляем mean_price_sq_by_disc – среднюю цену за квадратный метр в районе. Также убираем признак sub_area, так как он стал рудиментарным.

2.4.4. Параметры, относящиеся к транспортной системе.

В наборе данных есть признаки metro_km_walk, mkad_km, kremlin_km, metro_min_avto, metro_km_avto, metro_min_walk, которые так или иначе относятся к транспортной системе.

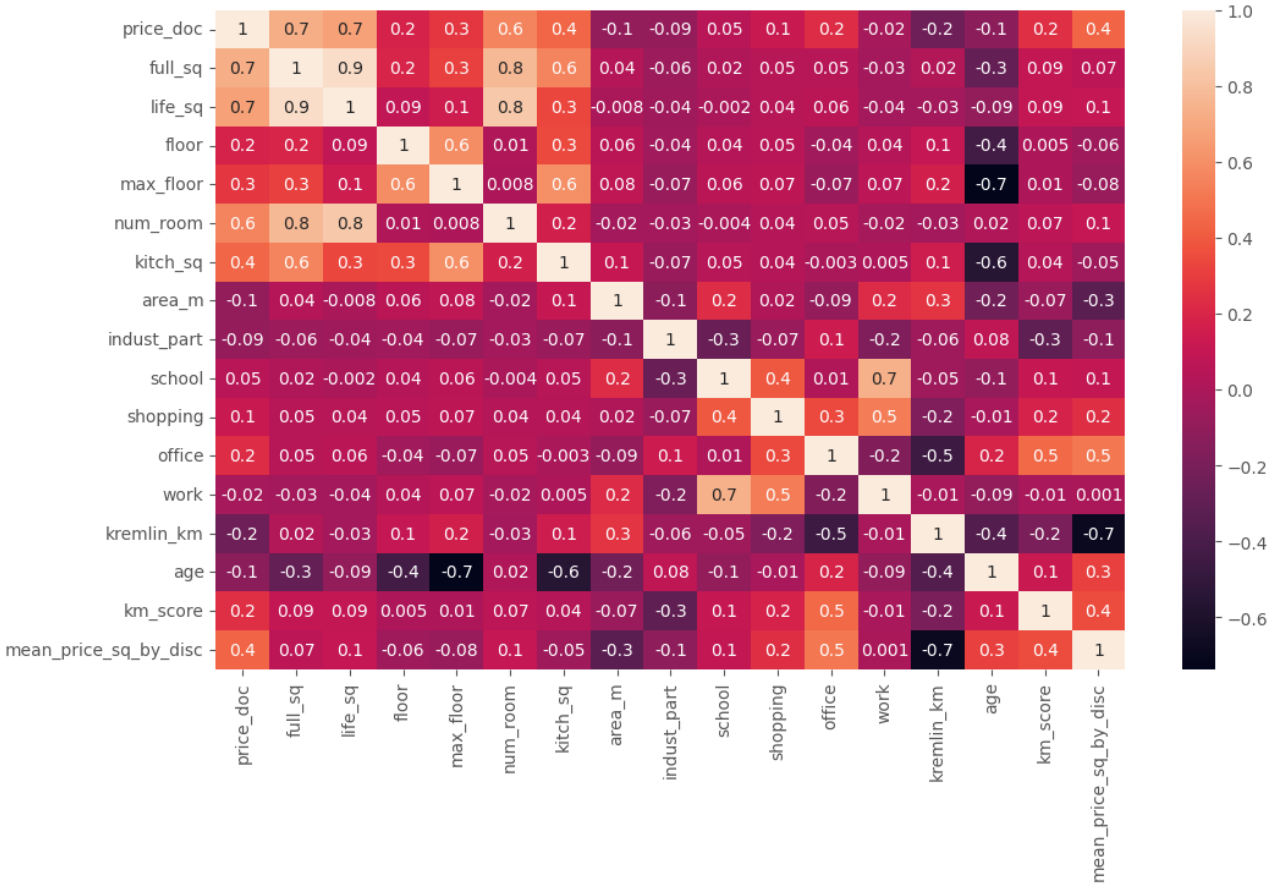
Посмотрим на матрицу корреляции:



Делаем вывод, что факторы сильно коррелируют друг с другом, поэтому стоит исключить большинство из них. При этом показатели расстояния от МКАД и Кремля чётко отражают местоположение квартиры.

2.4.5. Другие признаки.

Посмотрим на другие признаки и их корреляции:

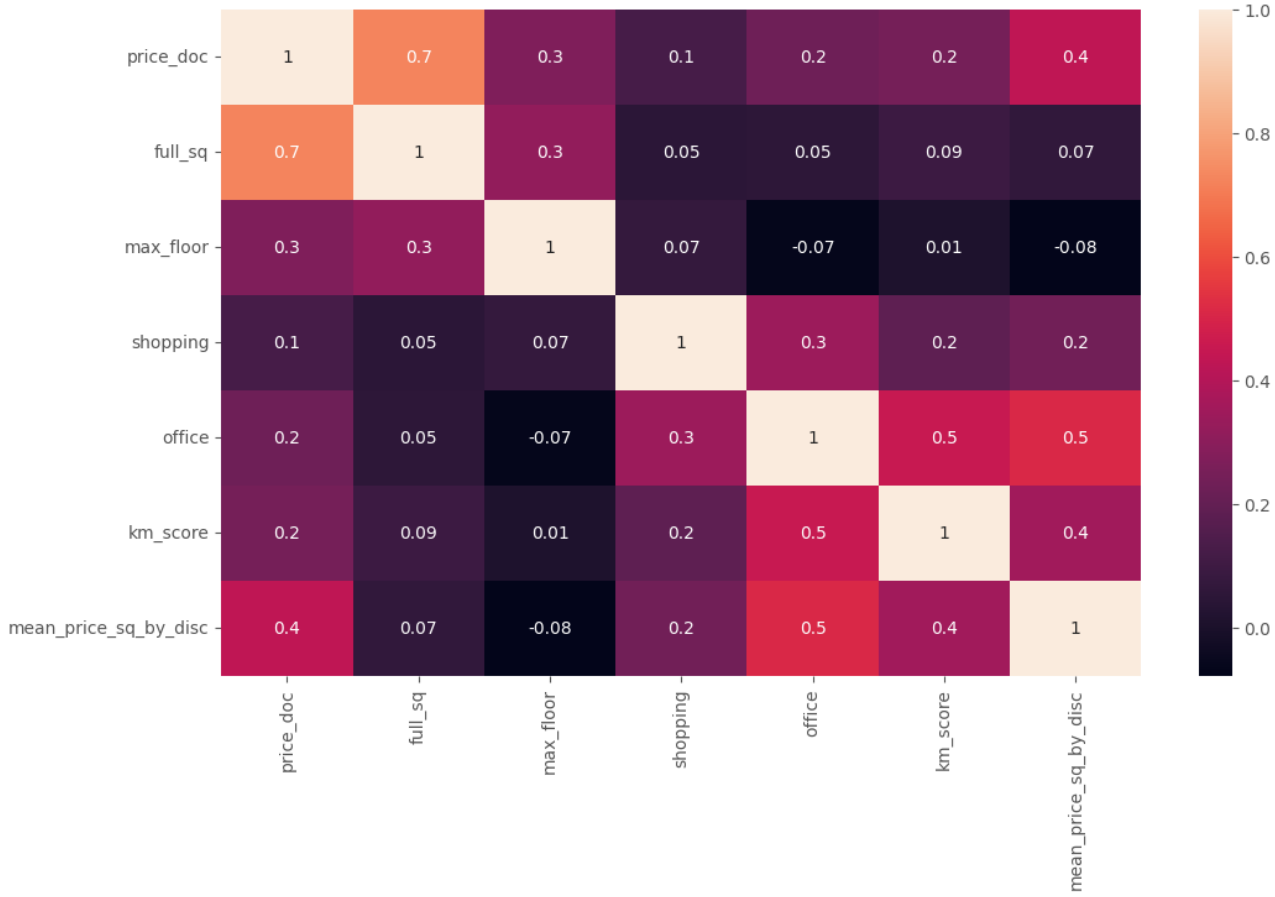


Признаки life_sq, num_room и kitch_sq создают сильную мультиколлинеарность с full_sq и не несут значимой смысловой нагрузки для прогнозирования цены, особенно учитывая, что методы расчёта этих параметров не всегда объективны.

Исключим оставшиеся транспортные признаки, так как в полном наборе данных было выявлено, что они демонстрирует сильную мультиколлинеарность. (буквально в предложении выше сказано то же самое).

Также мы исключаем этажи и возраст зданий, так как они создают мультиколлинеарность друг с другом и с параметром max_floors. При этом показатель max_floors может быть использован для связи цены с возрастом зданий.

Получим такую матрицу корреляции:

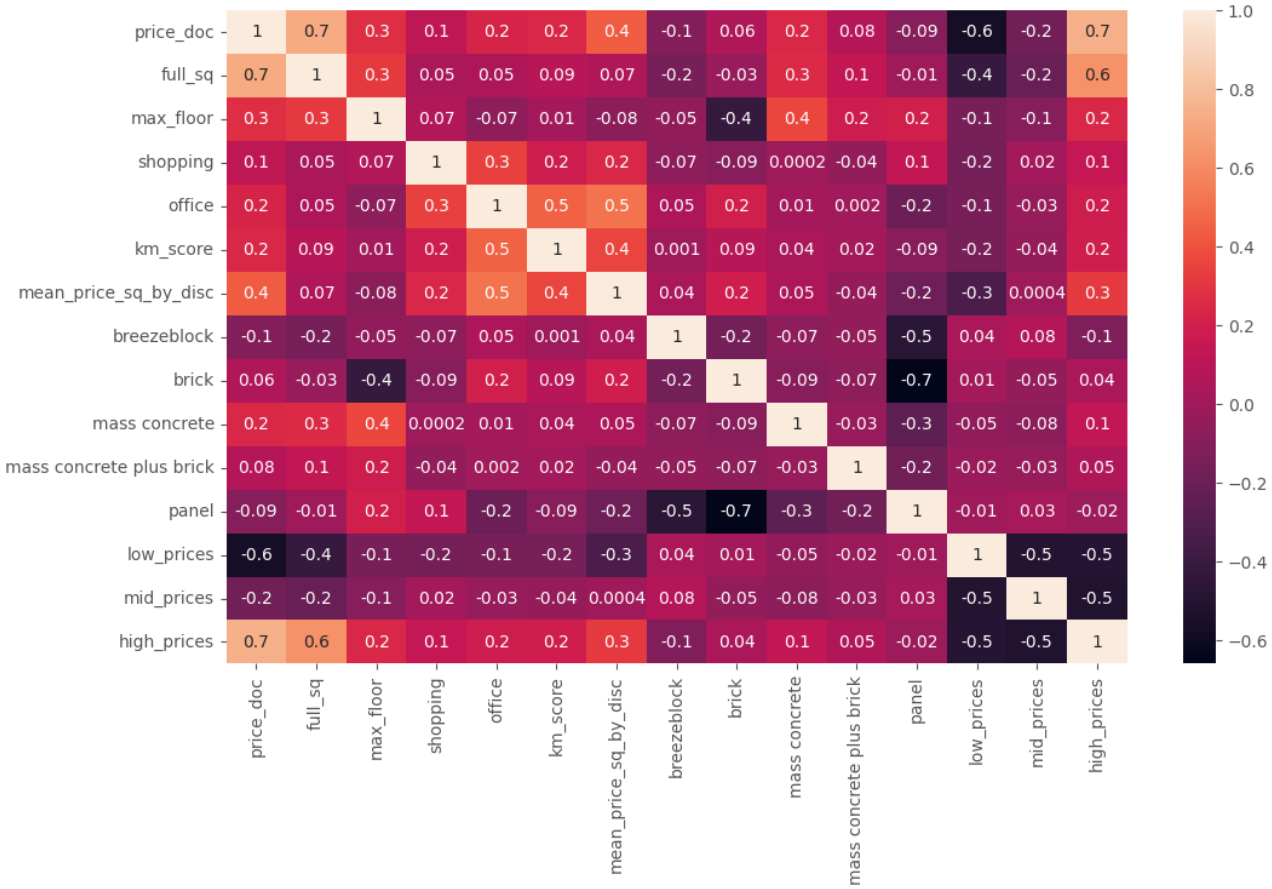


Далее сделаем one-hot encoding признака ‘material’ и создадим критерии для оценки стоимости квартиры:

- дешёвая — меньше 33-го процентиля,
- средняя — между 33-м и 66-м процентилями,
- дорогая — больше 66-го процентиля.

Построив box-plot для признаков (см. приложение), нетрудно заметить большое количество выбросов, но мы не будем их исключать, так как в этом случае потеряем много необходимых данных, которые не попадают в основную выборку.

В итоге получили проанализированные данные почти без выбросов и без сильной мультиколлинеарности.


price_doc 1 0.7 0.3 0.1 0.2 0.2 0.4 -0.1 0.06 0.2 0.08 -0.09 -0.6 -0.2 0.7
full_sq 0.7 1 0.3 0.05 0.05 0.09 0.07 -0.2 -0.03 0.3 0.1 -0.01 -0.4 -0.2 0.6
max_floor 0.3 0.3 1 0.07 -0.07 0.01 -0.08 -0.05 -0.4 0.4 0.2 0.2 -0.1 -0.1 0.2
shopping 0.1 0.05 0.07 1 0.3 0.2 0.2 -0.07 -0.09 0.0002 -0.04 0.1 -0.2 0.02 0.1
office 0.2 0.05 -0.07 0.3 1 0.5 0.5 0.05 0.2 0.01 0.002 -0.2 -0.1 -0.03 0.2
km_score 0.2 0.09 0.01 0.2 0.5 1 0.4 0.001 0.09 0.04 0.02 -0.09 -0.2 -0.04 0.2
mean_price_sq_by_disc 0.4 0.07 -0.08 0.2 0.5 0.4 1 0.04 0.2 0.05 -0.04 -0.2 -0.3 0.0004 0.3
breezeblock -0.1 -0.2 -0.05 -0.07 0.05 0.001 0.04 1 -0.2 -0.07 -0.05 -0.5 0.04 0.08 -0.1
brick 0.06 -0.03 -0.4 -0.09 0.2 0.09 0.2 -0.2 1 -0.09 -0.07 -0.7 0.01 -0.05 0.04
mass concrete 0.2 0.3 0.4 0.0002 0.01 0.04 0.05 -0.07 -0.09 1 -0.03 -0.3 -0.05 -0.08 0.1
mass concrete plus brick 0.08 0.1 0.2 -0.04 0.002 0.02 -0.04 -0.05 -0.07 -0.03 1 -0.2 -0.02 -0.03 0.05
panel -0.09 -0.01 0.2 0.1 -0.2 -0.09 -0.2 -0.5 -0.7 -0.3 -0.2 1 -0.01 0.03 -0.02
low_prices -0.6 -0.4 -0.1 -0.2 -0.1 -0.2 -0.3 0.04 0.01 -0.05 -0.02 -0.01 1 -0.5 -0.5
mid_prices -0.2 -0.2 -0.1 0.02 -0.03 -0.04 0.0004 0.08 -0.05 -0.08 -0.03 0.03 -0.5 1 -0.5
high_prices 0.7 0.6 0.2 0.1 0.2 0.2 0.3 -0.1 0.04 0.1 0.05 -0.02 -0.5 -0.5 1
price_doc
full_sq
max_floor
shopping
office
km_score
mean_price_sq_by_disc
breezeblock
brick
mass concrete
mass concrete plus brick
panel
low_prices
mid_prices
high_prices
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

2.5. Модели регрессии

У нас есть числовые признаки price_doc, full_sq, max_floor, shopping, office, km_score, mean_price_sq_by_disc, и номинативные признаки breezeblock, brick, mass concrete, mass concrete plus brick, panel, low_prices, mid_prices, high_prices.

Стандартизируем числовые признаки, т.е. приведем к значениям от –1 до 1.

	price_doc	full_sq	max_floor	shopping	office	km_score	mean_price_sq_by_disc
count	7770.00	7770.00	7770.00	7770.00	7770.00	7770.00	7770.00
mean	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00
std	1.00	1.00	1.00	1.00	1.00	1.00	1.00
min	-1.63	-2.10	-2.03	-1.22	-0.59	-3.73	-3.30
25%	-0.59	-0.72	-0.59	-0.72	-0.48	-0.37	-0.39
50%	-0.26	-0.34	-0.05	-0.21	-0.26	0.10	-0.12
75%	0.33	0.45	0.68	0.30	0.07	0.36	0.52
max	6.84	11.89	6.46	3.60	5.96	12.26	3.41

2.5.1. Многомерная линейная регрессия.

Разделим данные на обучающую и тестовую выборку 70 на 30 процентов соответственно.

Обучив модель, получим такие метрики:

- **MSE (Mean Squared Error) = 0.2670** — в случае стандартизированных данных это значение ошибки отражает, насколько хорошо модель предсказала значения, но важно понимать, что MSE для стандартизированных данных не имеет прямого смысла в единицах оригинальных данных, так как данные были приведены к единому масштабу. Таким образом, низкое значение MSE свидетельствует о хорошем качестве модели, но его интерпретация ограничена именно контекстом стандартизированных данных.
- **MAE (Mean Absolute Error) = 0.3162** — как и MSE, MAE для стандартизированных данных оценивает абсолютное отклонение предсказанных значений от реальных. Поскольку данные были стандартизированы, это значение

также представляет ошибку в стандартизированном масштабе. То есть, это значение показывает, на сколько единиц отклоняются предсказанные данные от реальных в стандартизированном масштабе.

- **r^2_score = 0.7385** — коэффициент детерминации остаётся одинаковым для стандартизированных данных, так как этот показатель зависит от объяснённой вариации в данных и не чувствителен к масштабированию данных. r^2 по-прежнему показывает, что модель объясняет около 74% вариации целевой переменной, что является хорошим результатом.

2.5.2. Алгоритм случайного леса.

Аналогично разделим данные на обучающую и тестовую выборку 70 на 30 процентов соответственно.

Обучив модель, получим такие метрики:

- **MSE = 0.1021** — среднеквадратическая ошибка, что указывает на хорошее качество предсказаний.
- **MAE = 0.1961** — средняя абсолютная ошибка, показывающая небольшие отклонения предсказанных значений от реальных.
- **r^2 = 0.8965** — коэффициент детерминации, указывающий, что модель объясняет около 89.65% вариации данных, что свидетельствует о высокой точности модели.

Воспользуемся **сеточным поиском гиперпараметров** (Grid Search), чтобы улучшить качество модели.

Лучшие параметры, найденные в результате поиска по сетке для модели случайного леса:

- **max_depth = None** — модель будет продолжать делить узлы до тех пор, пока все листья не будут чистыми (т.е. все объекты в узле будут иметь одинаковые метки).
- **max_features = None** — при разделении каждого узла будут использоваться все доступные признаки.

- **min_samples_leaf = 4** — минимальное количество образцов в листе дерева должно быть не менее 4. Это предотвращает создание узлов, состоящих из слишком маленьких групп данных.
- **min_samples_split = 10** — минимальное количество образцов, необходимых для разделения узла равно 10. Это ограничивает переобучение.
- **n_estimators = 300** — количество деревьев в случайном лесе, что указывает на более сложную модель с потенциально лучшей производительностью за счет использования большего количества деревьев.

Эти параметры показывают, что оптимальной конфигурацией для данной задачи является использование большого количества деревьев с глубиной, не ограниченной заранее, и с достаточно высокими значениями для параметров минимального количества образцов для разделения и для листа.

Обучив модель на тестовой выборке с подобранными параметрами, получим такие метрики:

- **$R^2 = 0.8988$** — Это значение коэффициента детерминации указывает на то, что модель объясняет около 89.88% вариации целевой переменной. Это незначительно лучше, чем на предыдущих этапах (где R^2 был 0.8965).
- **MSE = 0.0998** — Среднеквадратичная ошибка (MSE) также показывает улучшение по сравнению с предыдущими результатами (где MSE был 0.1021). Меньшее значение MSE говорит о том, что модель теперь предсказывает значения с меньшей ошибкой.

Далее используем **кросс-валидацию**, чтобы оценить стабильность модели и её способность обобщать на новых данных. Это позволяет нам избежать переобучения модели и оценить, насколько хорошо она будет работать на различных подвыборках данных.

После проведения кросс-валидации получены следующие результаты:

- **Средний $R^2 = 0.8408$** — это среднее значение коэффициента детерминации по всем фолдам, что показывает, что модель в среднем объясняет около 84.08% вариации целевой переменной на разных подвыборках данных.

- **Дисперсия = 0.0221** — низкая дисперсия указывает на то, что модель демонстрирует стабильные результаты, и её производительность не сильно зависит от конкретной подвыборки данных.

Сделаем вывод о стабильности модели: средний R^2 на кросс-валидации близок к метрике на тестовой выборке, что указывает на хорошую стабильность модели при её обобщении на разных подвыборках данных. Это подтверждается низкой дисперсией (0.0221), что означает, что модель даёт относительно постоянные результаты на разных подвыборках.

ЗАКЛЮЧЕНИЕ

В данной курсовой работе была рассмотрена одна из актуальных задач анализа данных – прогнозирование стоимости недвижимости, решение которой основывается на применении регрессионного анализа. Целью исследования являлось создание моделей, пригодных для прогнозирования стоимости недвижимости в Москве на основе различных факторов. Данная цель была достигнута путем выполнения следующих задач:

1. Изучение теоретических основ регрессионного анализа.
2. Сбор и обработка статистических данных для анализа.
3. Построение и обучение регрессионных моделей.
4. Оптимизация параметров моделей для повышения их точности.
5. Проведение оценки адекватности построенных моделей.

В рамках исследования были рассмотрены два метода регрессии: многомерная линейная регрессия и регрессия на основе случайного леса.

Для работы с данными и построения моделей использовались библиотеки Python, такие как **NumPy**, **Pandas**, **Scikit-learn**, **Statsmodels** и другие.

В процессе работы выполнена предварительная обработка данных: стандартизация, проверка на наличие выбросов и анализ корреляций между признаками.

Для оценки моделей использовались метрики $\mathbf{R^2}$, среднеквадратичная ошибка (MSE) и средняя абсолютная ошибка (MAE). Было выявлено, что модели регрессии демонстрируют приемлемую точность.

В ходе сравнения метрик моделей установлено, что регрессия на основе случайного леса показывает более высокую точность на тестовой выборке, достигая $\mathbf{R^2 = 0.8988}$, что свидетельствует о её превосходстве над многомерной линейной регрессией на данном наборе данных.

Результаты исследования могут быть использованы для прогнозирования стоимости недвижимости в зависимости от различных факторов, таких как удаленность от центра города, близость к метро, инфраструктура района и

других. Разработанные модели могут найти применение при анализе рынка недвижимости и в поддержке принятия инвестиционных решений.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Б. А. Севастьянов “Курс теории вероятностей и математической статистики”
2. Документация к python библиотеке scikit-learn. - https://scikit-learn.org/stable/
3. Яндекс хендбук - учебник по машинному обучению. - https://education.yandex.ru/handbook/ml
4. Alex Smola, S.V.N. Vishwanathan “Introduction to Machine Learning”
5. John O. Rawlings, Sastry G. Pantula, David A. Dickey “Applied Regression Analysis: A Research Tool, Second Edition”