Оглавление

1 Общая постановка задачи

1.1 Формулировка прикладной проблемы

В настоящей работе исследуется зависимость цен частных домов в США от характеристик как самого дома, так и участка земли, на котором он расположен.

Актуальность данного исследования подтверждается тем, что большой процент американцев предпочитает проживать в загородных домах. Соответственно, задача о покупке жилья загородом по приемлемому соотношению цены к качеству является востребованной к анализу.

В данной работе рассматривается *задача определения оптимальной стоимости загородного дома с заданными параметрами, а именно, многоэтажного дома без подвала возрастом не более 30 лет с жилой площадью не менее 2021 квадратных футов, площадью участка около 8500 квадратных футов, который не был на реконструкции.*

1.2 Потенциальные потребители решения; задачи, которые они смогут решать, используя полученные результаты

Результаты, полученные в ходе исследования, могут использоваться потенциальными покупателями частных домов в США, участниками американского рынка недвижимости. Потребители смогут оценить примерную стоимость дома, удовлетворяющую заданным критериям, и определить свои возможности на приобретение жилья.

1.3 Основные гипотезы, которые планируется проверить в рамках решения задачи

В таблице 1 представлены характеристики домов, используемые для анализа их стоимости.

№	Характеристика объекта	Название переменной	Шкала измерения	Роль переменной
1	Жилая площадь, кв. футы	sqft_living	Относительная	Независимая
2	Площадь участка, кв. футы	sqft_lot	Относительная	Независимая
3	Одноэтажный	floor	Номинальная (дихотомическая)	Независимая
4	Наличие подвала	basement	Номинальная (дихотомическая)	Независимая
5	Год постройки	year_built	Относительная	Независимая

6	Реконструкция	renovated	Номинальная (дихотомическая)	Независимая
7	Цена дома*, т. долларов	price	Относительная	Зависимая

Таблица 1. Описание факторов, учтенных в анализе

* под ценой дома понимается общая стоимость участка земли и самого дома.

Сформулируем гипотезы о статистической взаимосвязи зависимой переменной и независимыми:

1. *С увеличением возраста дома его цена падает, при этом скорость падения цены отремонтированных домов ниже, чем неотремонтированных.*

Основа этой гипотезы заключается в том, что старые дома подвержены большему риску, к примеру, вероятны проблемы с канализацией или отоплением, как следствие, потенциальные расходы покупателя на содержание дома увеличиваются. В свою очередь реконструированные дома подвержены меньшему риску, что добавляет им стоимости.

2. *По мере увеличения площади участка цена на дом значительно увеличивается до некоторого значения, а потом перестает заметно расти.*

Такое предположение обусловлено высокими ценами на землю, а значит, дом на большом участке ориентирован на более состоятельных потребителей. При этом из-за повышенных налогов на землю цена не может неограниченно расти, поэтому предполагается существование некоторого значения площади участка, после которого темп роста цены на дом сильно падает.

3. *Многоэтажные дома стоят дороже одноэтажных, причем наличие подвала также влияет на рост цены дома, но в меньшей степени*.

Данная гипотеза основана на том, что многоэтажные дома обычно имеют больший метраж, чем одноэтажные, а значит, и стоят дороже. Наличие подвала также увеличивает стоимость жилья (за счет увеличения метража), но в меньшей степени.

1.4 Основные источники данных

Основные данные были взяты с **Kaggle.com** из датасета **House price prediction**, содержащего информацию об основных характеристиках частных домов в США.

2 Предварительный анализ собранных данных

2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

В этой части отчета будет выполнен общий анализ используемых данных, будут построены гистограммы, описывающие зависимые и независимые переменные. Следует отметить, что предварительный анализ исходной выборки данных требовал некоторой коррекции, а именно, удаления тех записей, которые априори можно было считать недействительными или ошибочными. К подобным элементам выборки относятся записи, содержащие пропуски, нулевые значения (например, нулевые цены) или неприемлемые величины (например, очевидно огромные значения цен или площадей участков). Подобные записи были сочтены нами невалидными, и, следовательно, были исключены из дальнейшего рассмотрения. В результате начальный объем выборки сократился до 4400 элементов.

2.1.1 Анализ количественных переменных

Для дальнейшего анализа входящих в наше исследование количественных переменных «Жилая площадь» (sqft_living), «Площадь участка» (sqft_lot), «Год постройки» (year_built), «Цена» (price) приведем таблицу основных статистик (таблица 2).

Переменная	Среднее	Медиана	Минимум	Максимум
sqft_living	2117.8	1970.0	370.0	9640.0
sqft_lot	10505.0	7560.0	638.0	99916.0
year_built	1970.8	1976.0	1900.0	2014.0
price	549.41	465.0	149.5	4668.0
Переменная	**Ст. отклонение**	**Вариация**	**Асимметрия**	**Эксцесс**
sqft_living	907.79	0.42865	1.2766	3.4
sqft_lot	11196.0	1.0657	3.3619	13.888
year_built	29.86	0.015151	-0.50525	-0.66615
price	352.9	0.64232	3.1847	18.442
Переменная	**5% перцентиль**	**95% перцентиль**	**Межквартильный размах**	**Пропущенные набл.**
sqft_living	960.0	3809.5	1135.3	0
sqft_lot	1613.0	35398.0	5537.5	0
year_built	1912.0	2009.0	46.0	0
price	219.03	1183.8	325.83	0

Таблица 2. Описательная статистика количественных переменных

Анализ количественной переменной «Жилая площадь»

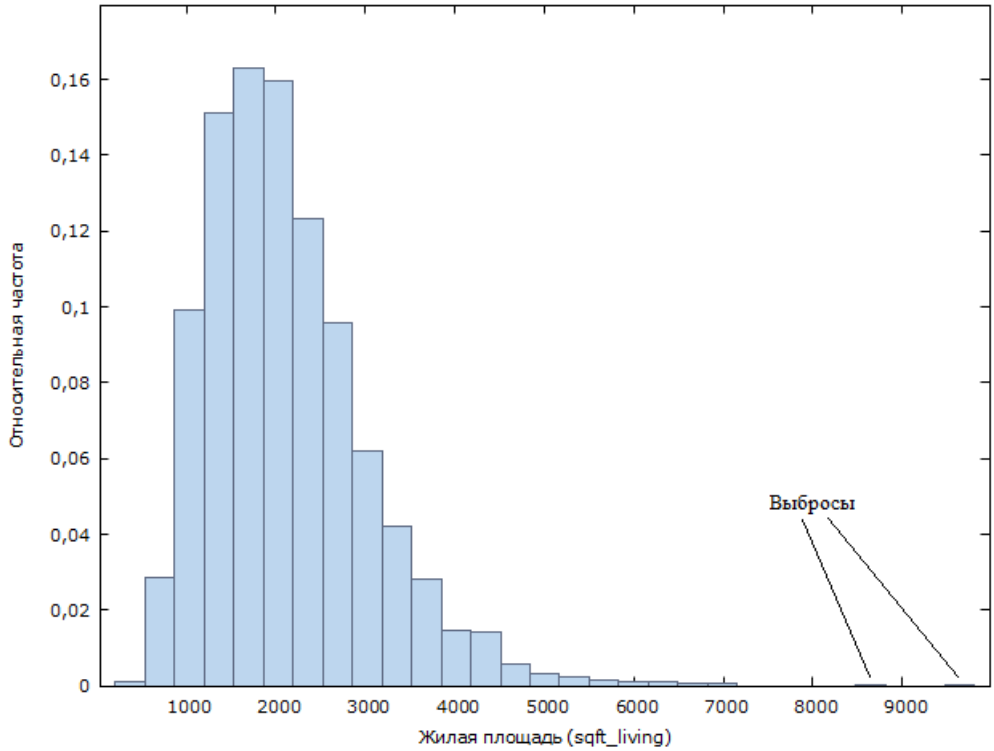

Рисунок 1. Жилая площадь

На рисунке 1 приведена гистограмма, описывающая распределение частот переменной «Жилая площадь» (sqft_living).

Анализируя гистограмму (см. рисунок 1) и беря во внимание таблицу описательных статистик количественных переменных (см. таблицу 2), следует сделать вывод о том, что в распределении данной переменной присутствует асимметрия вправо (подтверждается тем, что среднее (2117.8) больше медианы (1970.0), а также тем, что коэффициент асимметрии (1.2766) положительный). Это объясняется таким фактом, что основная часть домов имеет среднюю (более-менее стандартную) жилую площадь, но при этом существуют и дома с намного большим метражом, что полностью соответствует действительности.

Из представленной выше диаграммы видно, что в распределении этой переменной наблюдается заметный «пик» по сравнению с нормальным распределением (подтверждается тем, что коэффициент эксцесса (3.4) положительный), значит, можно говорить об островершинности. Из этого следует неоднородность в выборке, которую мы наблюдаем, – большая часть значений лежит в диапазоне стандартных площадей дома, что в целом естественно.

Полимодальности в выборке не наблюдается, так как в распределении виден только один «пик».

Тест на нормальное распределение хи-квадрат (926.710) показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза о нормальности распределения переменной может быть отвергнута.

Для выявления выбросов в выборке воспользуемся правилом «трех сигм», которое гласит, что значения, отстоящие от среднего (2117.8) более чем на утроенное стандартное отклонение (3*907.79), могут быть интерпретированы как выбросы. Для исследуемой переменной выбросами считаются элементы выборки, не попавшие в промежуток (0, 4841.17). Их число равняется 48 измерениям, что составляет 1.09% выборки.

Анализ количественной переменной «Площадь участка»

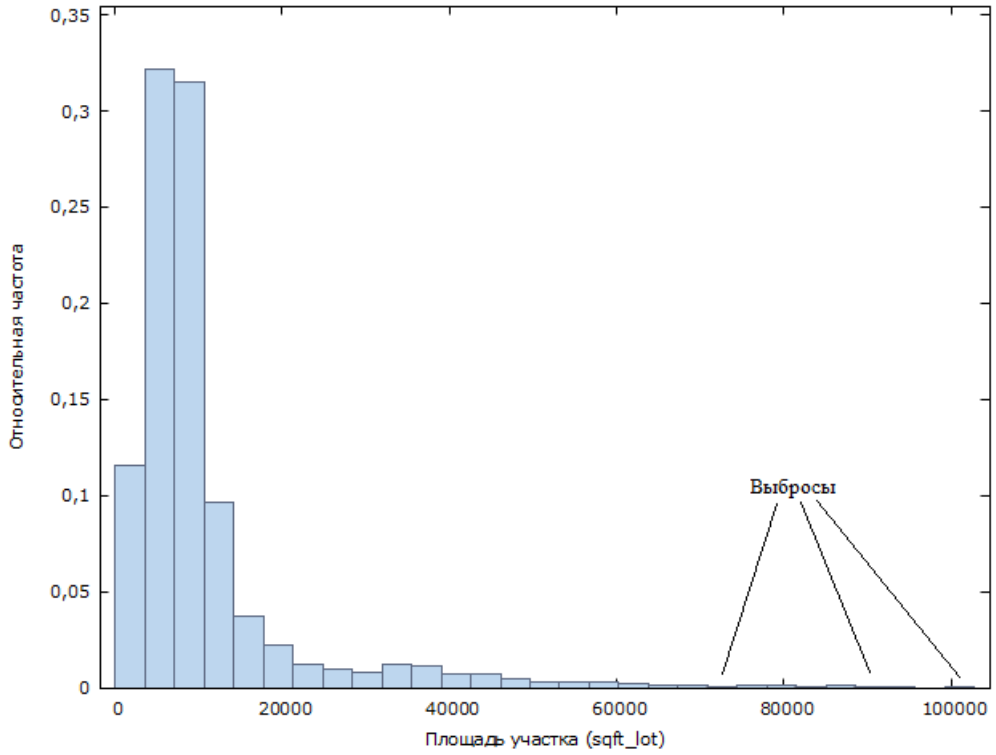

Рисунок 2. Площадь участка

На рисунке 2 приведена гистограмма, описывающая распределение частот переменной «Площадь участка» (sqft_lot).

Анализируя гистограмму (см. рисунок 2) и беря во внимание таблицу описательных статистик количественных переменных (см. таблицу 2), следует сделать вывод о том, что в распределении данной переменной присутствует асимметрия вправо (подтверждается тем, что среднее (10505.0) больше медианы (7560.0), а также тем, что коэффициент асимметрии (3.3619) положительный). Это объясняется таким фактом, что основная часть домов имеет небольшую/среднюю (более-менее стандартную) площадь участка, но при этом существуют и дома с намного большим метражом участка, что полностью соответствует действительности.

Из представленной выше диаграммы видно, что в распределении этой переменной наблюдается ярко выраженный «пик» по сравнению с нормальным распределением (подтверждается тем, что коэффициент эксцесса (13.888) положительный), значит, можно говорить об островершинности. Из этого следует неоднородность в выборке,

которую мы наблюдаем, – большая часть значений лежит в диапазоне стандартных площадей участка, что в целом отражает реальность.

Полимодальности в выборке не наблюдается, так как в распределении виден только один «пик».

Тест на нормальное распределение хи-квадрат (17012.859) показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза о нормальности распределения переменной может быть отвергнута.

Для выявления выбросов в выборке воспользуемся правилом «трех сигм», которое гласит, что значения, отстоящие от среднего (10505.0) более чем на утроенное стандартное отклонение (3*11196.0), могут быть интерпретированы как выбросы. Для исследуемой переменной выбросами считаются элементы выборки, не попавшие в промежуток (0, 44093.0). Их число равняется 118 измерениям, что составляет 2.68% выборки.

Анализ количественной переменной «Год постройки»

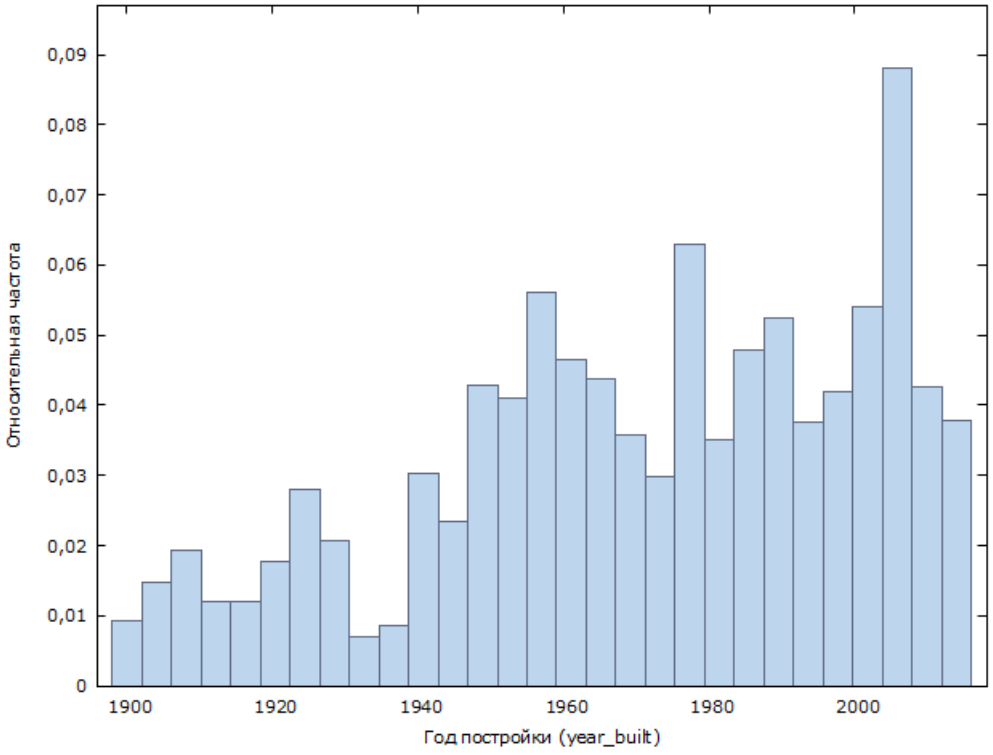

Рисунок 3. Год постройки

На рисунке 3 приведена гистограмма, описывающая распределение частот переменной «Год постройки» (year_built).

Анализируя гистограмму (см. рисунок 3) и беря во внимание таблицу описательных статистик количественных переменных (см. таблицу 2), следует сделать вывод о том, что в распределении данной переменной присутствует асимметрия влево (подтверждается тем, что среднее (1970.8) меньше медианы (1976.0), а также тем, что коэффициент асимметрии (-0.50525) отрицательный). Это объясняется таким фактом,

что основная часть домов была построена во второй половине 20-го – начале 21-го века, но при этом в выборке присутствуют и более старые дома, построенные в начале 20-го века. Эта картина соответствует действительности, поскольку обычно очень старые дома сносят, чтобы построить новые, а не продают, но тем не менее все же встречаются дома, которые были построены в первой половине 20-го века.

Из представленной выше диаграммы видно, что в распределении этой переменной как таковой «пик» не наблюдается (подтверждается тем, что коэффициент эксцесса (-0.66615) отрицательный), значит, не можем говорить об островершинности. Из этого следует, что наблюдаемая выборка достаточно однородная.

Как таковой полимодальности в выборке не наблюдается.

Тест на нормальное распределение хи-квадрат (664.325) показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза о нормальности распределения переменной может быть отвергнута.

Для выявления выбросов в выборке воспользуемся правилом «трех сигм», которое гласит, что значения, отстоящие от среднего (1970.8) более чем на утроенное стандартное отклонение (3*29.86), могут быть интерпретированы как выбросы. Для исследуемой переменной выбросами считаются элементы выборки, не попавшие в промежуток (1881.22, 2060.38). Их число равняется 0 измерениям и составляет 0% выборки.

Анализ количественной переменной «Цена»

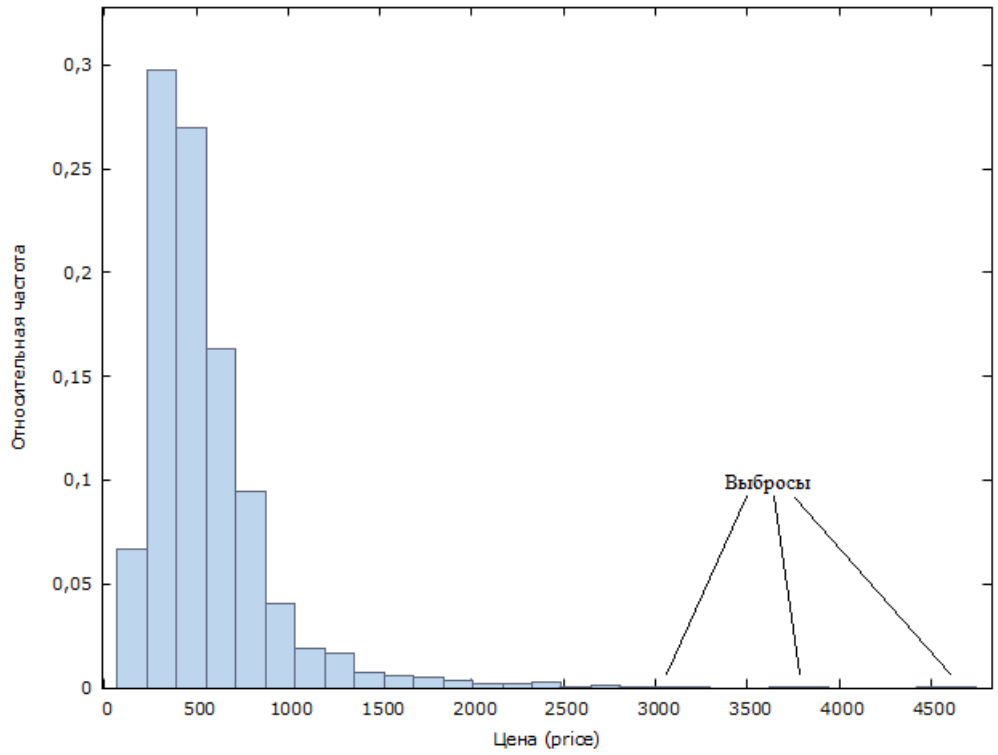

Рисунок 4. Цена

На рисунке 4 приведена гистограмма, описывающая распределение частот переменной «Цена» (price).

Анализируя гистограмму (см. рисунок 4) и беря во внимание таблицу описательных статистик количественных переменных (см. таблицу 2), следует сделать вывод о том, что в распределении данной переменной присутствует асимметрия вправо (подтверждается тем, что среднее (549.41) больше медианы (465.0), а также тем, что коэффициент асимметрии (3.1847) положительный). Это объясняется таким фактом, что основная часть домов имеет стандартную цену, но при этом существуют и дома, которые стоят намного дороже. Это вполне естественно, поскольку в реальности встречаются дома, которые продаются по более высокой цене из-за внешних причин (например, очень хорошее расположение дома или широкая известность предыдущих владельцев).

Из представленной выше диаграммы видно, что в распределении этой переменной наблюдается очень заметный «пик» по сравнению с нормальным распределением (подтверждается тем, что коэффициент эксцесса (18.442) положительный), значит, можно говорить об островершинности. Из этого следует неоднородность в выборке, которую мы наблюдаем, – большая часть значений лежит в диапазоне стандартных цен на дом, что вполне схоже с действительностью.

Полимодальности в выборке не наблюдается, так как в распределении виден только один «пик».

Тест на нормальное распределение хи-квадрат (5813.037) показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза о нормальности распределения переменной может быть отвергнута.

Для выявления выбросов в выборке воспользуемся правилом «трех сигм», которое гласит, что значения, отстоящие от среднего (549.41) более чем на утроенное стандартное отклонение (3* 352.9), могут быть интерпретированы как выбросы. Для исследуемой переменной выбросами считаются элементы выборки, не попавшие в промежуток (0, 1608.11). Их число равняется 89 измерениям, что составляет 2.02% выборки.

Дальнейший анализ этих элементов выборки показал, что целесообразно считать их выбросами из-за завышенной цены, поскольку среднее жилых площадей (4410) этих измерений более чем в 2 раза превышает среднее жилых площадей (2118) по всей выборке из 4400 измерений, а также среднее метражей участков (15584) этих элементов в 1.5 раза больше среднего по метражам участков (10505) всей выборки, а эти показатели естественно отражаются на ценах домов. Более того около 85% домов, помеченных как выбросы, являются многоэтажными, в то время как в изначальной выборке таковыми являются чуть более 50% домов, что, предположительно, также влияет на увеличение их стоимости.

2.1.2 Анализ качественных переменных

Ниже на рисунках 5–7 приведены столбчатые диаграммы, отражающие количество измерений с разными уровнями для качественных переменных «Одноэтажный» (floor), «Подвал» (basement), «Реконструкция» (renovated) и «Состояние» (condition) соответственно.

Анализ качественной переменной «Одноэтажный»

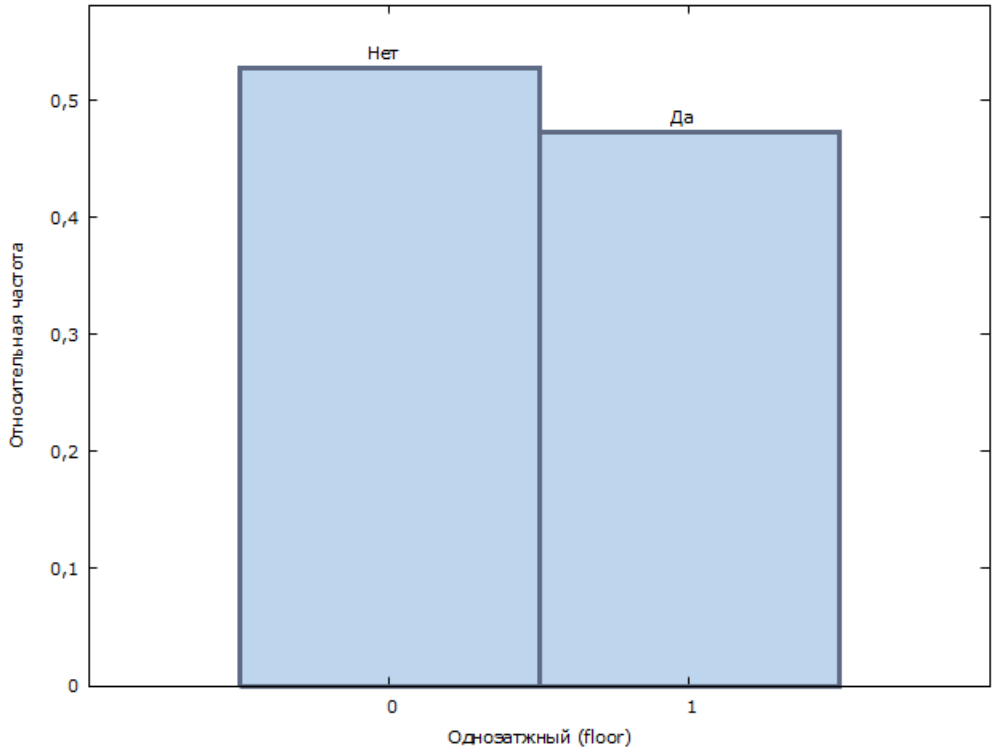

Рисунок 5. Количество этажей

По распределению частот количества этажей (см. рисунок 5) выборка является сбалансированной. Видно, что одноэтажных домов примерно 46% от общего числа строений, многоэтажных же больше половины. Дома высотою в 3 этажа и более – достаточно редкое явление в США, значит, большую часть категории «многоэтажные дома» составляют строения с 2 этажами. Отсюда следует, что одноэтажных и двухэтажных домов приблизительно одинаковое количество, что в целом отражает действительность.

Уровней с долей менее 5% не было выявлено, необходимости в укрупнении нет.

Анализ качественной переменной «Наличие подвала»

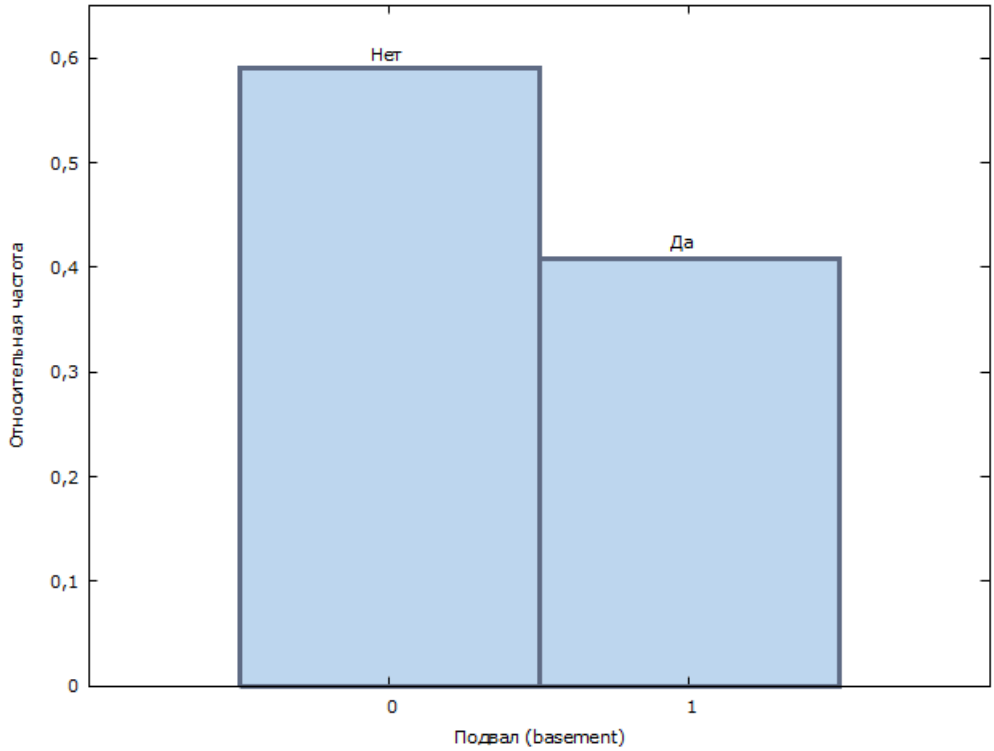

Рисунок 6. Наличие подвала

Из построенной гистограммы видно, что примерно у 60% домов нет подвалов (см. рисунок 6). Полученная картина в целом соответствует реальности, поскольку несколько большая доля людей предпочитает не переплачивать за подвал, который не является жилой площадью, но при этом добавляет стоимости дому за счет увеличения метража. Предполагается, что такой расклад учитывается при строительстве домов. Тем не менее в какой-то степени подвал может доставлять удобства жильцам как дополнительное помещение, которое можно использовать как склад, поэтому дома с подвалами составляют немалую долю от всех домов, что подтверждается диаграммой.

Уровней с долей менее 5% не было выявлено, необходимости в укрупнении нет.

Анализ качественной переменной «Реконструкция»

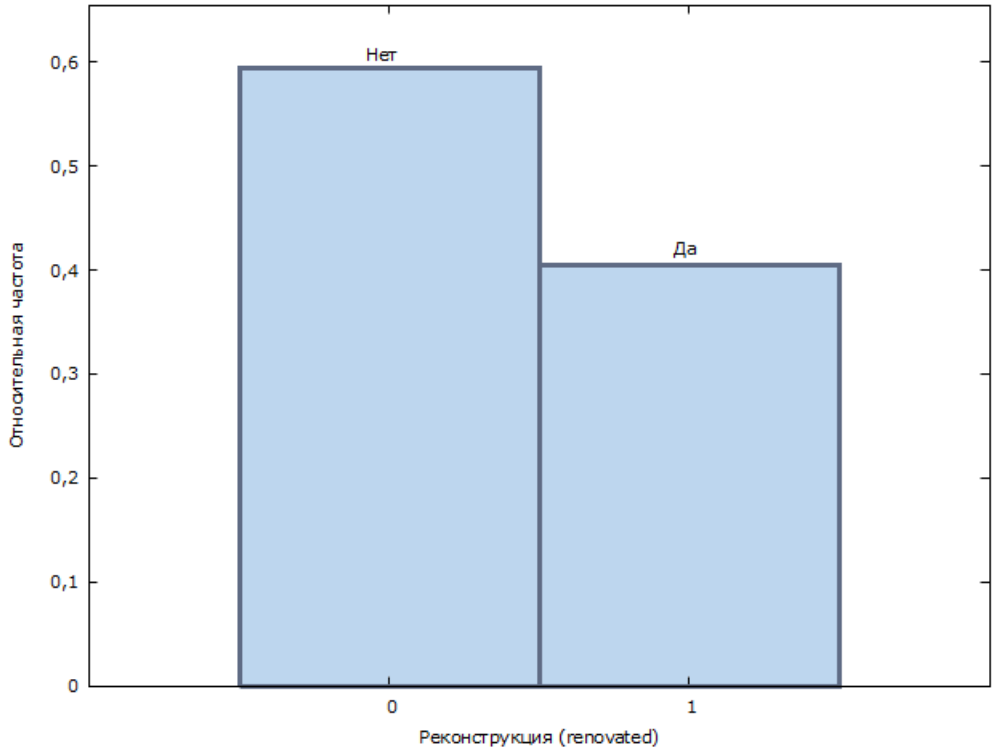

Рисунок 7. Реконструкция

Около 40% домов были реконструированы (см. рисунок 7). Полученная цифра соответствует реальности, так как достаточно большое количество домов (см. рисунок 3) были построены еще в первой половине 20 века. Люди предпочитают жить в более новых домах, поэтому старые дома необходимо ремонтировать. Более того самые старые дома, очевидно, строились по устаревшим технологиям, а значит у таких домов существует риск возникновения проблем с канализацией, отоплением и прочими необходимыми для комфортного проживания условиями.

Уровней с долей менее 5% не было выявлено, необходимости в укрупнении нет.

2.1.3 Анализ репрезентативности выборки

По окончании анализа количественных и качественных переменных можно сделать вывод о том, что выборка репрезентативна, поскольку преимущественно отражает американскую действительность рынка домов.

2.2 Анализ статистической связи

2.2.1 Графический анализ пары «числовая зависимая переменная – качественная независимая переменная»

Чтобы проанализировать пары «числовая зависимая переменная – качественная независимая переменная» для каждой качественной переменной, построим

категорированные диаграммы Бокса-Уискера и приведем результаты выполнения теста Краскела-Уолиса.

Анализ пары «Цена» – «Одноэтажный»

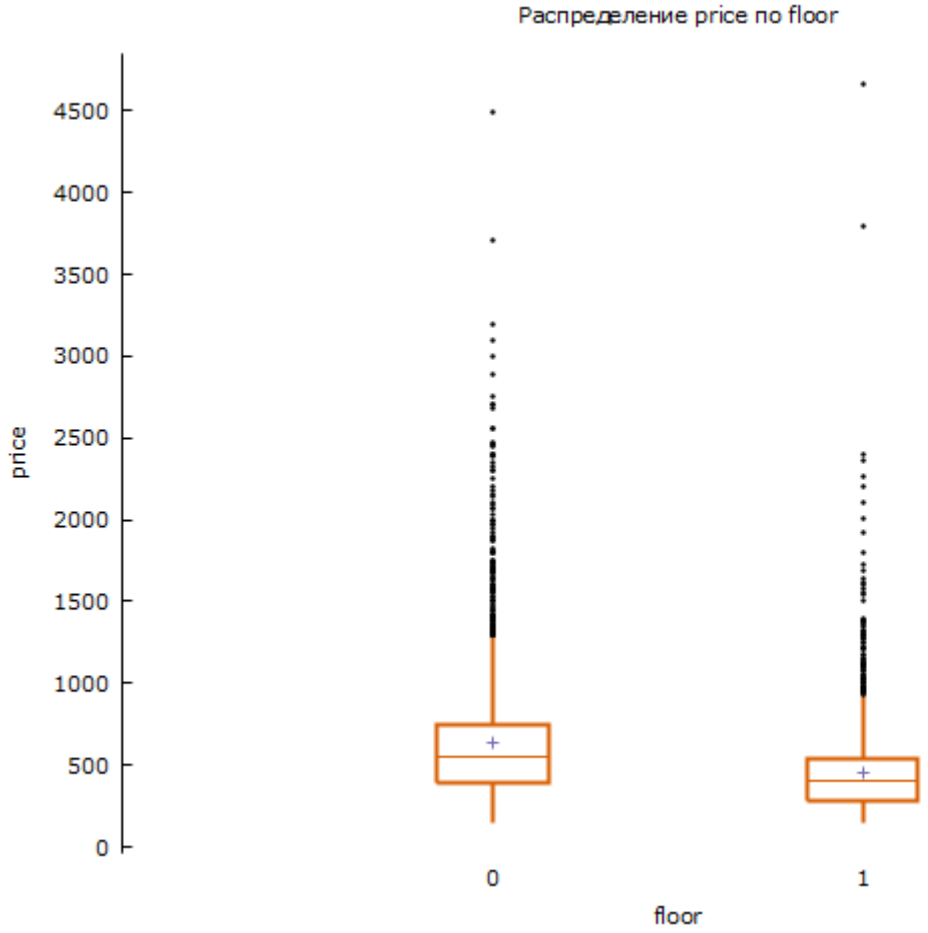

Рисунок 8. Зависимость цены от количества этажей (0 – многоэтажный, 1 – одноэтажный)

Анализируя представленную на рисунке 8 коробчатую диаграмму, можно сделать вывод о том, что как медиана цен многоэтажных домов превышает медиану цен одноэтажных, так и разброс цен для первых больше, чем разброс цен для вторых. Таким образом, предварительно можно утверждать, что существуют значимые различия между стоимостью многоэтажных и одноэтажных домов. Медиана цен многоэтажных домов соответствует верхнему квартилю цен на одноэтажные дома, при этом медиана цен на одноэтажные дома соответствует нижнему квартилю цен многоэтажных домов. Полученные результаты соответствуют действительности, поскольку справедливо предполагать, что в среднем многоэтажные дома стоят дороже, и что диапазон их цен шире, чем у одноэтажных.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 463.6825, df = 1, p-value = 7.59505e-103), который показал,

что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Цена» – «Наличие подвала»

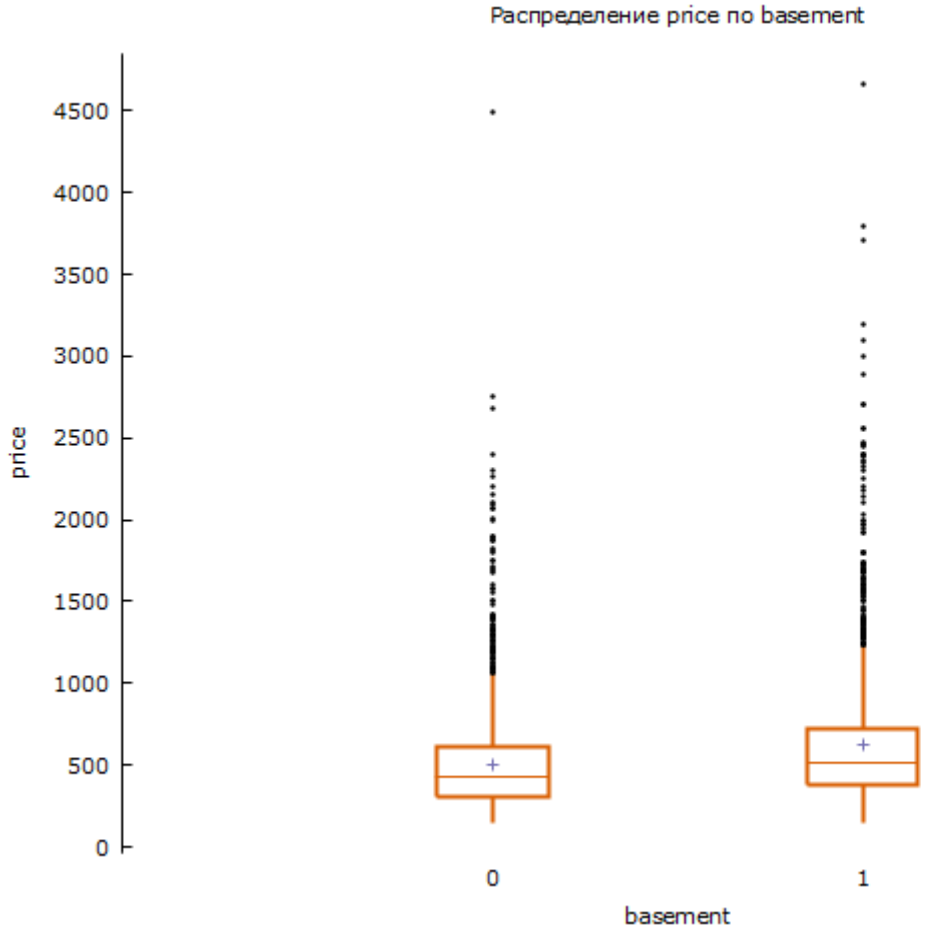

Рисунок 9. Зависимость цены от наличия подвала (0 – нет подвала, 1 – есть подвал)

Анализируя представленную на рисунке 9 коробчатую диаграмму, можно сделать вывод о том, что как медиана цен домов с подвалом превышает медиану цен домов без подвала, так и разброс цен для первых несколько больше, чем разброс цен для вторых. Таким образом, предварительно можно утверждать, что существуют значимые различия между стоимостью домов с подвалами и без. Медиана цен домов без подвала несколько превышает нижний квартиль цен на дома с подвалом. Полученные результаты соответствуют действительности, поскольку справедливо предполагать, что в среднем дома с подвалом стоят дороже, и что диапазон их цен несколько шире, чем у домов без подвала.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 153.6788, df = 1, p-value = 2.7222e-035), который показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Цена» – «Реконструкция»

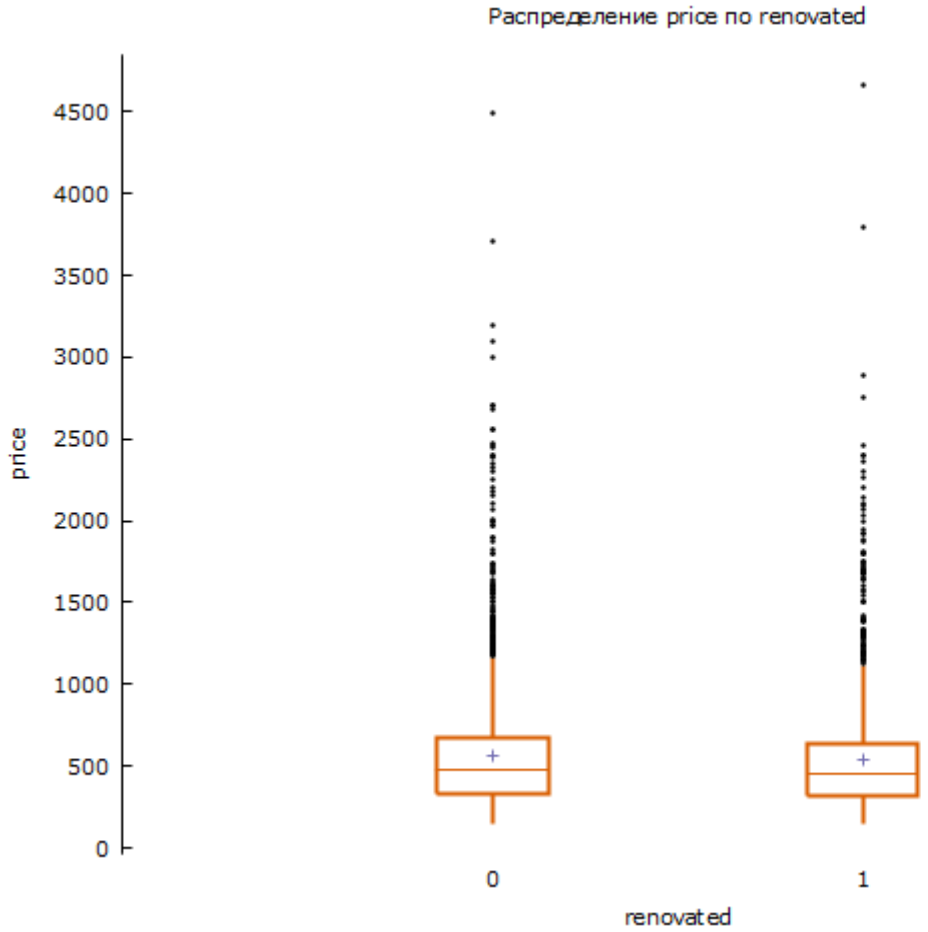

Рисунок 10. Зависимость цены от реконструкции (0 – не было реконструкции, 1 – была реконструкция)

Анализируя представленную на рисунке 10 коробчатую диаграмму, можно сделать вывод о том, что как медиана цен неотремонтированных домов незначительно превышает медиану цен домов с ремонтом, так и разброс цен для первых немного больше, чем разброс цен для вторых. Таким образом, можно предполагать, что значимых различий между стоимостью домов с ремонтом и без может не быть. Полученные результаты можно интерпретировать, опираясь на то, что в среднем дома без ремонта обычно новее и будут стоить несколько дороже.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 13.9838, df = 1, p-value = 0.000184392), который показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

2.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная»

Чтобы проанализировать пары «числовая зависимая переменная – числовая независимая переменная» для каждой количественной переменной, построим диаграммы рассеивания и приведем значения коэффициентов корреляции Пирсона, Спирмена и тау Кендалла.

Анализ пары «Цена» – «Жилая площадь»

Рисунок 11. Зависимость цены от жилой площади

Анализируя представленный на рисунке 11 график, можно сказать, что между исследуемыми переменными существует хорошо заметная прямая зависимость, то есть с увеличением жилой площади растет цена. Присутствуют очевидные выбросы – измерения либо со слишком выделяющейся большой жилой площадью, либо со значительно завышенной ценой. Полученные результаты отражают реальность, поскольку метраж дома является одним из факторов, определяющих его стоимость.

Для формальной проверки гипотезы о наличии связи были посчитаны значения коэффициентов корреляции Пирсона, Спирмена и тау Кендалла, результаты приведены в таблице 3.

	Пирсона	Спирмена	Тау Кендалла
Коэффициент корреляции	0.68867549	0.64223332	0.46065418
Значимость	0.0000	0.0000	0.0000

Таблица 3. Коэффициенты корреляции для пары «Цена» – «Жилая площадь»

Коэффициенты корреляции Пирсона, Спирмена и тау Кендалла являются значимыми на уровне 5% (0.05), таким образом, мы получаем, что все коэффициенты значимы. Значения коэффициентов говорят о том, что существует умеренная положительная статистическая связь.

Анализ пары «Цена» – «Площадь участка»

Рисунок 12. Зависимость цены от площади участка

Анализируя представленный на рисунке 12 график, можно сказать, что между исследуемыми переменными усматривается незначительная прямая зависимость, то есть в целом с увеличением площади участка цена растет. Присутствуют очевидные выбросы – измерения либо со слишком выделяющимся большим метражом, либо с сильно завышенной ценой. На рассеивающей диаграмме видно заметное скопление измерений в пределах метража участка от 0 до 20000, поэтому предлагается разбить выборку на две подвыборки, чтобы провести более детальное исследование.

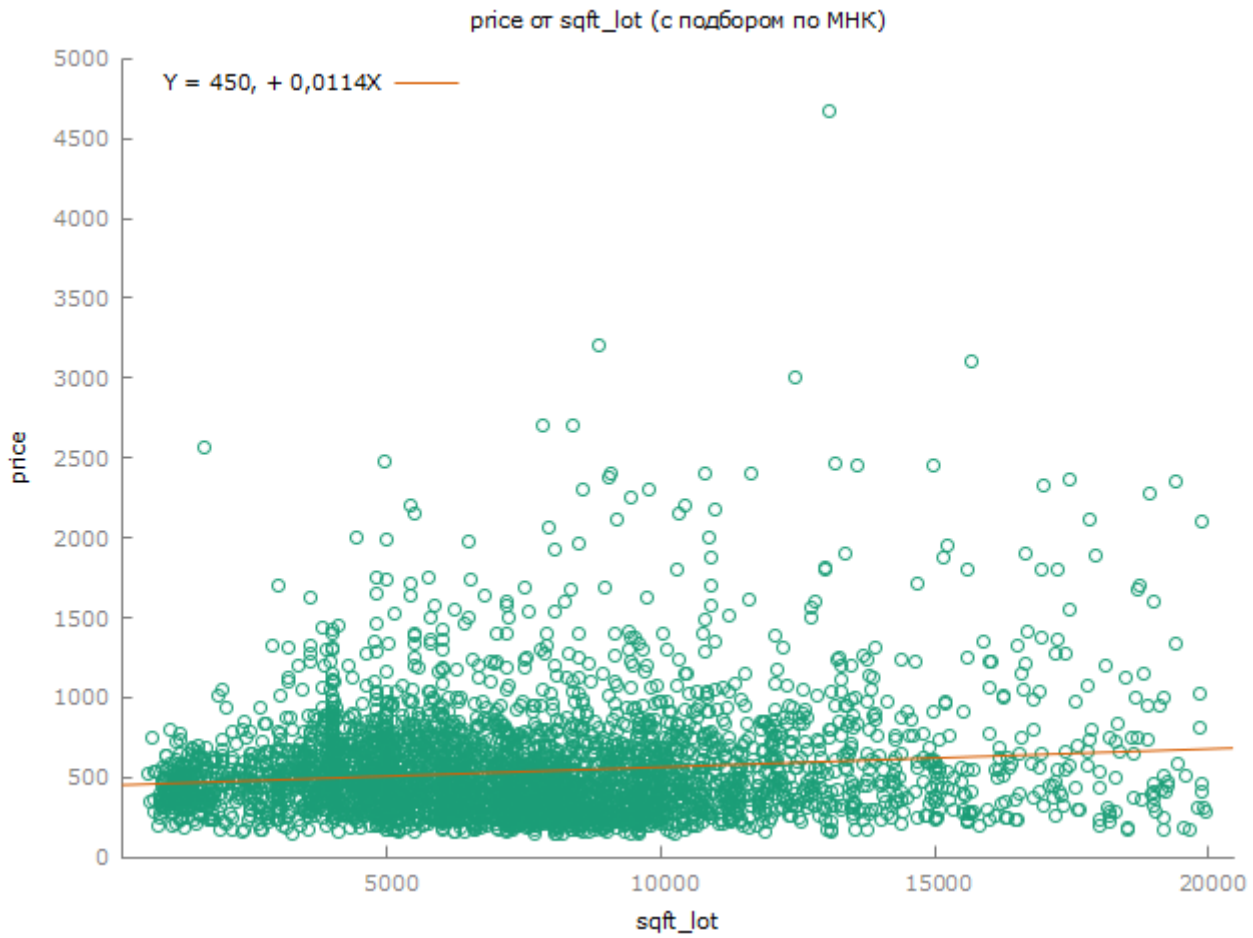

Рисунок 13. Зависимость цены от площади участка (до 20000 кв. футов)

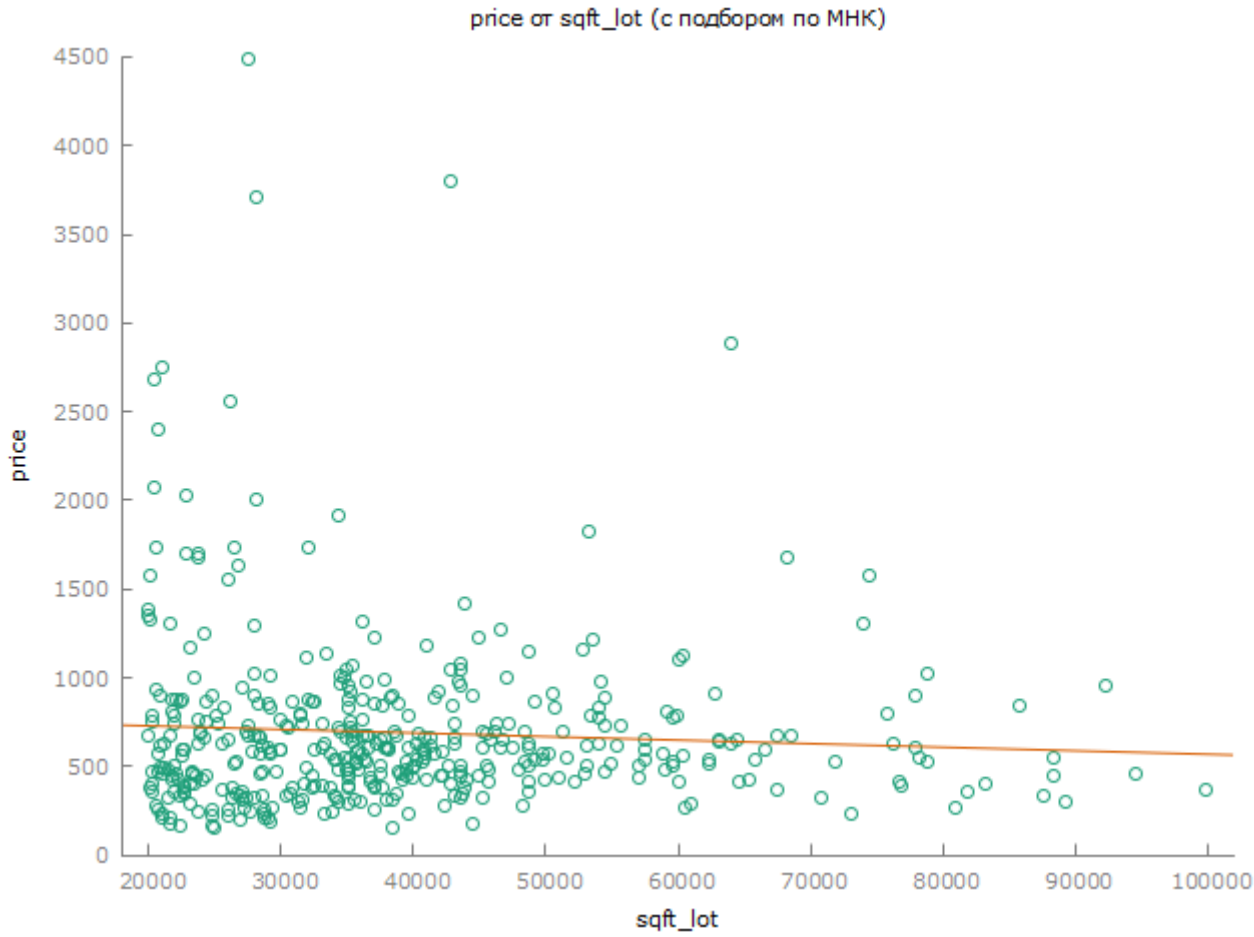

Рисунок 14. Зависимость цены от площади участка (от 20000 кв. футов)

Анализируя представленный на рисунке 13 график, можно увидеть более заметную прямую зависимость между переменными, в то время как на рисунке 14 представлен график, на котором видна обратная зависимость между переменными. Значит, можно предполагать, что по мере увеличения площади участка цена сначала растет до некоторого значения, но после может даже начать уменьшаться.

Для формальной проверки гипотезы о наличии связи были посчитаны значения коэффициентов корреляции Пирсона, Спирмена и тау Кендалла, результаты приведены в таблице 4.

	Пирсона	Спирмена	Тау Кендалла
Коэффициент корреляции	0.13520236	0.07299452	0.04651497
Значимость	0.0000	0.0000	0.0000

Таблица 4. Коэффициенты корреляции для пары «Цена» – «Площадь участка»

Коэффициенты корреляции Пирсона, Спирмена и тау Кендалла являются значимыми на уровне 5% (0.05), таким образом, мы получаем, что все коэффициенты значимы. Значения коэффициентов говорят о том, что если положительная статистическая связь и существует, то она слабая.

Анализ пары «Цена» – «Год постройки»

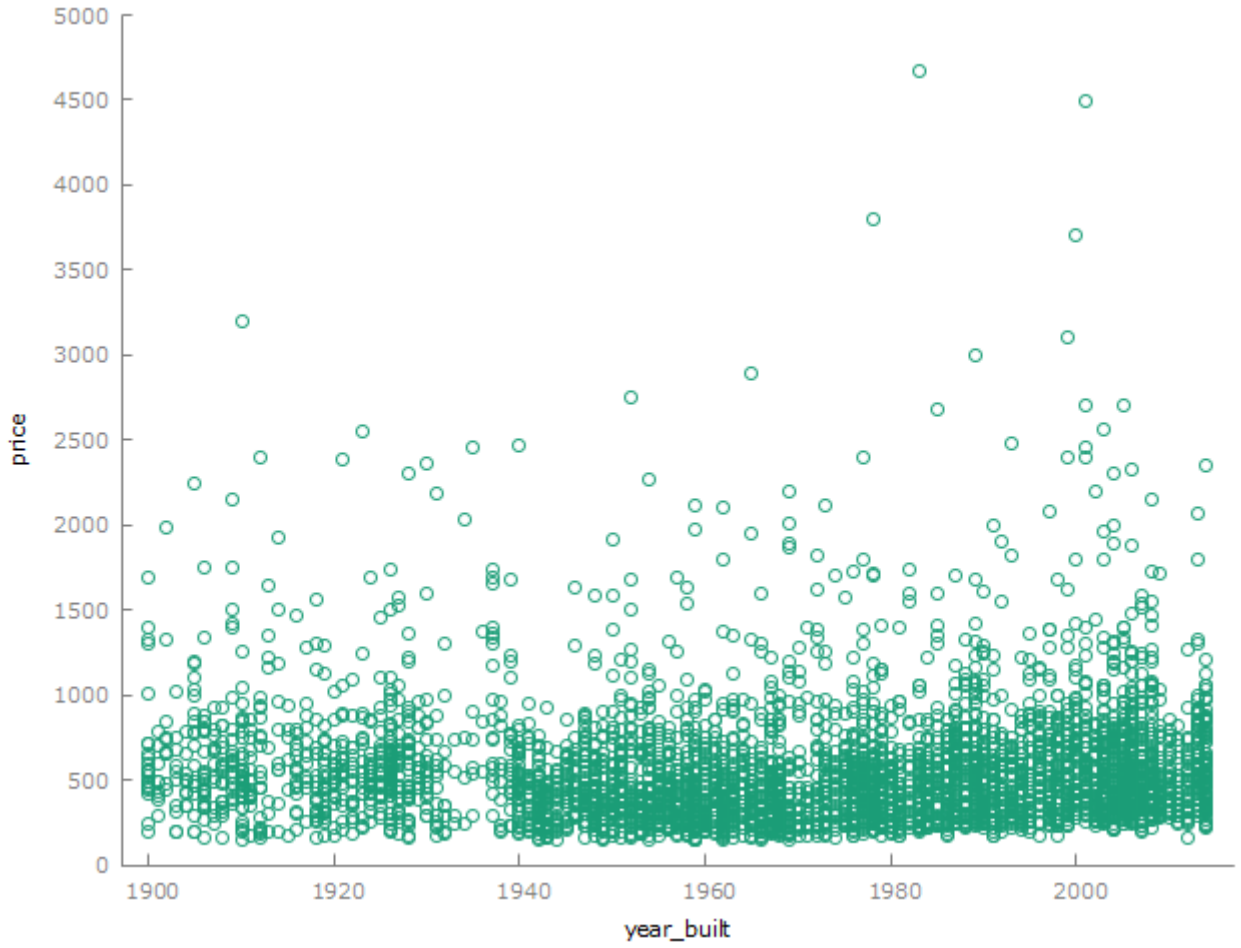

Рисунок 15. Зависимость цены от года постройки

Анализируя представленный на рисунке 15 график, можно сказать, что на первый взгляд между исследуемыми переменными не усматривается зависимость, то есть год постройки дома не влияет на его цену. Присутствуют очевидные выбросы – измерения с сильно завышенной ценой.

Для формальной проверки гипотезы о наличии связи были посчитаны значения коэффициентов корреляции Пирсона, Спирмена и тау Кендалла, результаты приведены в таблице 5.

	Пирсона	Спирмена	Тау Кендалла
Коэффициент корреляции	0.01201347	0.06257605	0.04300820
Значимость	0.4256	0.0000	0.0000

Таблица 5. Коэффициенты корреляции для пары «Цена» – «Год постройки»

Коэффициенты корреляции Пирсона, Спирмена и тау Кендалла являются значимыми на уровне 5% (0.05), таким образом, мы получаем, что коэффициент Пирсона незначим, а коэффициенты Спирмена и тау Кендалла значимы. Значения коэффициентов говорят о том, что если положительная статистическая связь и существует, то она крайне слабая.

2.2.3 Анализ наличия корреляции между независимыми переменными

Анализ наличия корреляции между независимыми качественными переменными. Чтобы проанализировать пары «качественная независимая переменная – качественная независимая переменная», построим таблицы сопряженности и приведем значения статистики хи-квадрат и коэффициента Крамера.

Анализ пары «Подвал» – «Одноэтажный»

Подвал / Одноэтажный	Нет (0)	Да (1)	Всего
Нет (0)	1667	656	2323
Да (1)	934	1143	2077
Всего	2601	1799	4400

Таблица 6. Таблица сопряженности для пары «Подвал» – «Одноэтажный»

Для формальной проверки гипотезы о наличии связи между исследуемыми переменными были посчитаны статистика хи-квадрат (325.668, 1 ст. свободы, p-значение = 8.43829e-073) и коэффициент Крамера (0.272), которые показали, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута. Значение коэффициента Крамера (0.272) говорит о наличии слабой связи.

Анализ пары «Реконструкция» – «Одноэтажный»

Реконструкция / Одноэтажный	Нет (0)	Да (1)	Всего
Нет (0)	1601	722	2323
Да (1)	1016	1061	2077
Всего	2617	1783	4400

Таблица 7. Таблица сопряженности для пары «Реконструкция» – «Одноэтажный»

Для формальной проверки гипотезы о наличии связи между исследуемыми переменными были посчитаны статистика хи-квадрат (182.039, 1 ст. свободы, p-значение = 1.73868e-041) и коэффициент Крамера (0.203), которые показали, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута. Значение коэффициента Крамера (0.203) говорит о наличии слабой связи.

Анализ пары «Реконструкция» – «Подвал»

Реконструкция / Подвал	Нет (0)	Да (1)	Всего
Нет (0)	1615	986	2601
Да (1)	1002	797	1799
Всего	2617	1783	4400

Таблица 8. Таблица сопряженности для пары «Реконструкция» – «Подвал»

Для формальной проверки гипотезы о наличии связи между исследуемыми переменными были посчитаны статистика хи-квадрат (18.0385, 1 ст. свободы, p-значение = 2.16485e-005) и коэффициент Крамера (0.064), которые показали, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута. Значение коэффициента Крамера (0.064) говорит о наличии крайне слабой связи.

Анализ наличия корреляции между независимыми числовыми переменными. Чтобы проанализировать пары «числовая независимая переменная – числовая независимая переменная», построим диаграммы рассеивания и приведем значения коэффициентов корреляции Пирсона, Спирмена и тау Кендалла.

Анализ пары «Площадь участка» – «Жилая площадь»

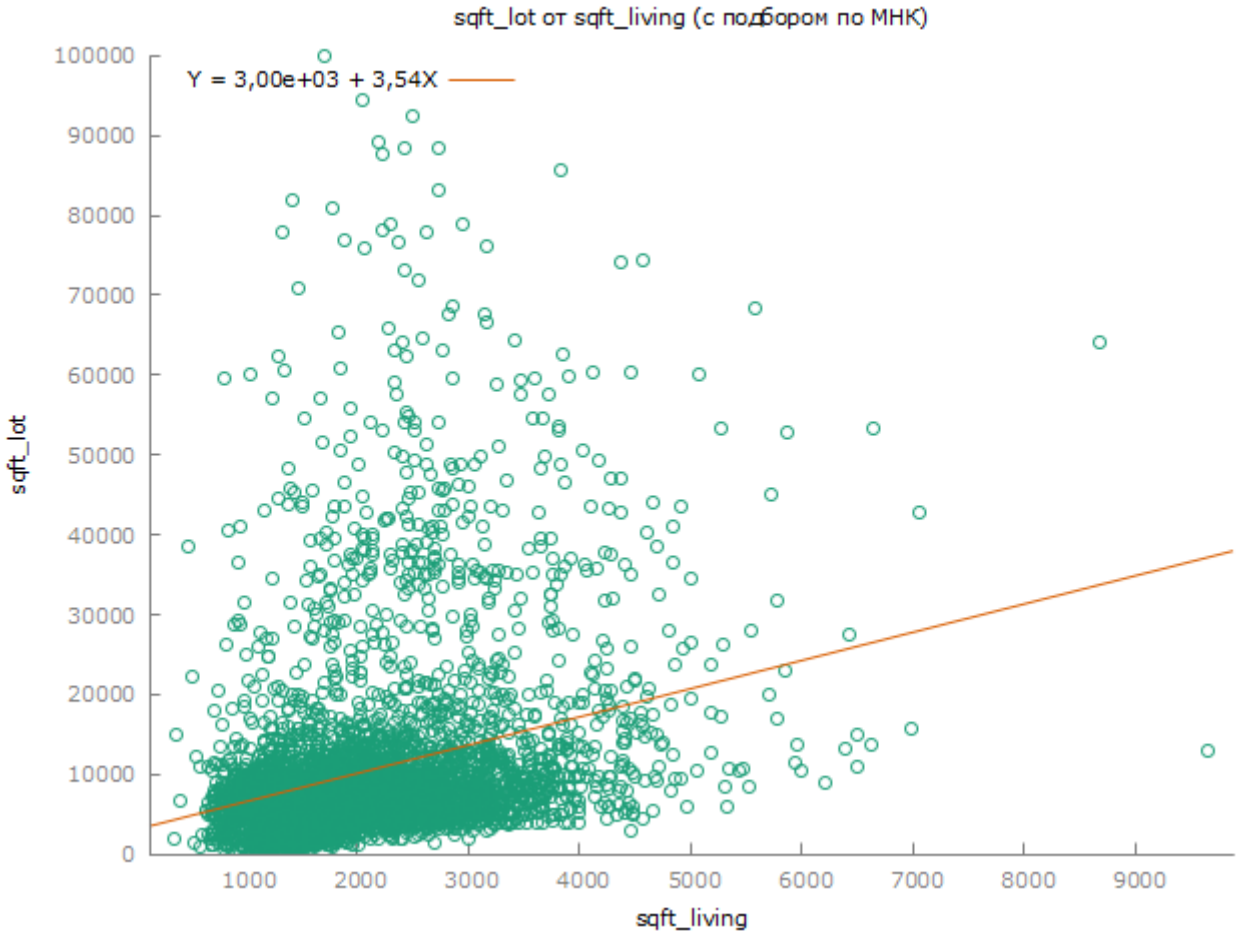

Рисунок 16. Зависимость площади участка от жилой площади

Анализируя представленный на рисунке 16 график, можно сказать, что между исследуемыми переменными существует заметная прямая зависимость, то есть с увеличением жилой площади дома растет величина площади участка.

Для формальной проверки гипотезы о наличии связи были посчитаны значения коэффициентов корреляции Пирсона, Спирмена и тау Кендалла, результаты приведены в таблице 9.

	Пирсона	Спирмена	Тау Кендалла
Коэффициент корреляции	0.28720510	0.31450656	0.21400309
Значимость	0.0000	0.0000	0.0000

Таблица 9. Коэффициенты корреляции для пары «Площадь участка» – «Жилая площадь»

Коэффициенты корреляции Пирсона, Спирмена и тау Кендалла являются значимыми на уровне 5% (0.05), таким образом, мы получаем, что все коэффициенты значимы. Значения коэффициентов говорят о том, что существует слабая положительная статистическая связь.

Анализ пары «Жилая площадь» – «Год постройки»

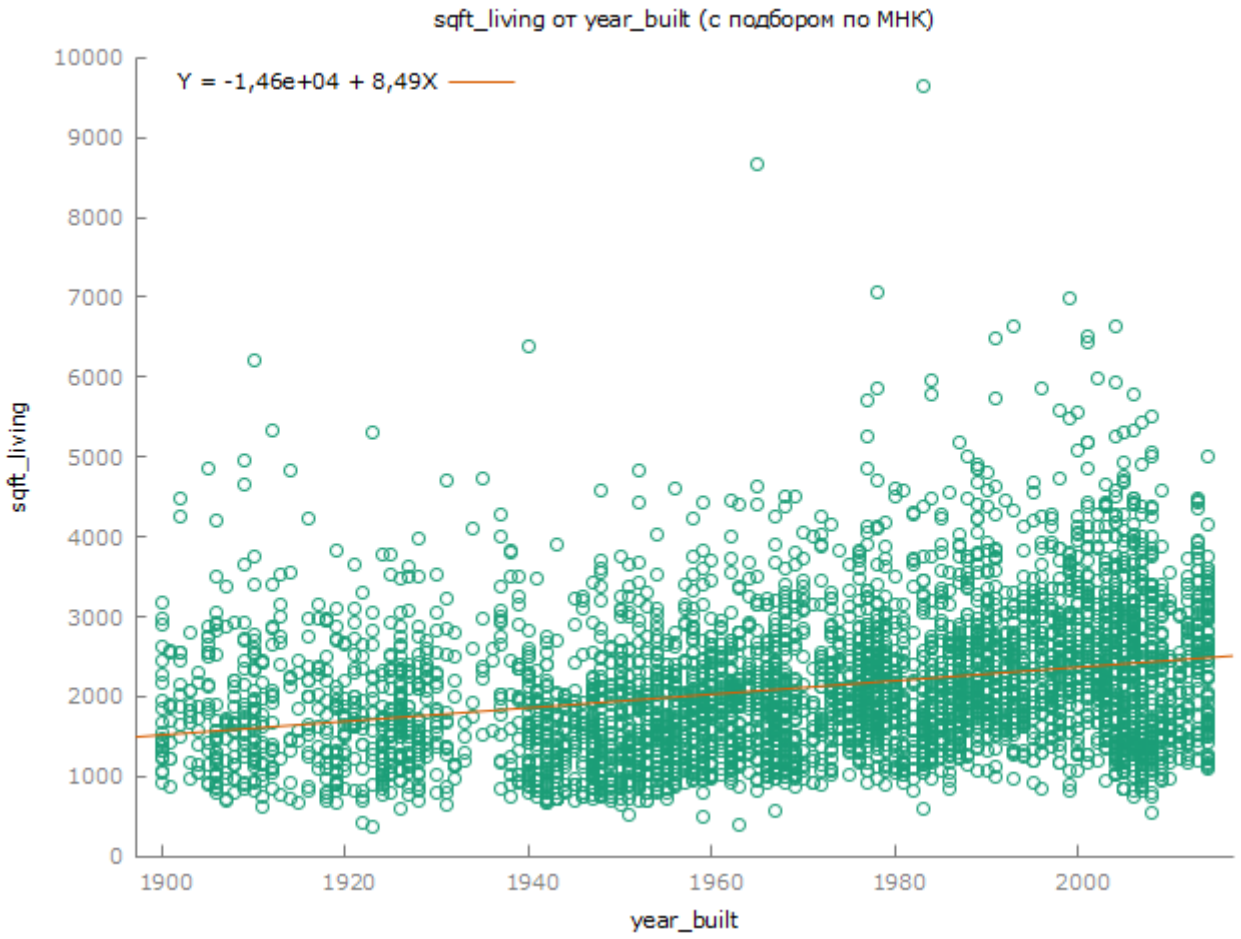

Рисунок 17. Зависимость жилой площади от года постройки

Анализируя представленный на рисунке 17 график, можно предположить, что между исследуемыми переменными может существовать слабая прямая зависимость, то есть в целом с увеличением года постройки дома (то есть с уменьшением возраста дома) растет значение жилой площади.

Для формальной проверки гипотезы о наличии связи были посчитаны значения коэффициентов корреляции Пирсона, Спирмена и тау Кендалла, результаты приведены в таблице 10.

	Пирсона	Спирмена	Тау Кендалла
Коэффициент корреляции	0.27916637	0.30908608	0.21419765
Значимость	0.0000	0.0000	0.0000

Таблица 10. Коэффициенты корреляции для пары «Жилая площадь» – «Год постройки»

Коэффициенты корреляции Пирсона, Спирмена и тау Кендалла являются значимыми на уровне 5% (0.05), таким образом, мы получаем, что все коэффициенты значимы. Значения коэффициентов говорят о том, что существует слабая положительная статистическая связь.

Анализ пары «Площадь участка» – «Год постройки»

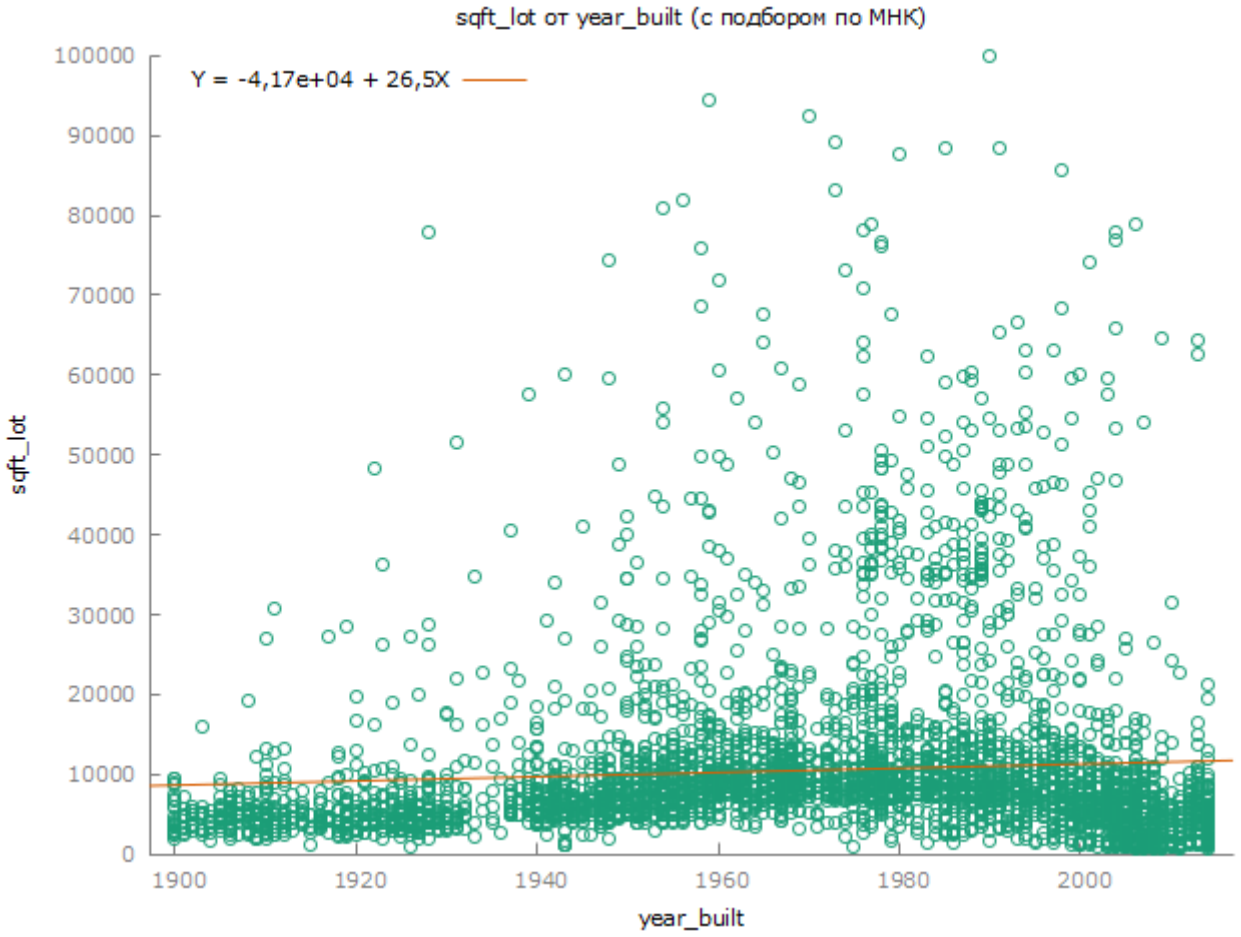

Рисунок 18. Зависимость площади участка от года постройки

Анализируя представленный на рисунке 18 график, можно сказать, что между исследуемыми переменными может существовать прямая зависимость, то есть в целом с увеличением года постройки дома (то есть с уменьшением возраста дома) растет величина площади участка.

Для формальной проверки гипотезы о наличии связи были посчитаны значения коэффициентов корреляции Пирсона, Спирмена и тау Кендалла, результаты приведены в таблице 11.

	Пирсона	Спирмена	Тау Кендалла
Коэффициент корреляции	0.07071717	-0.02781094	0.00007499
Значимость	0.0000	0.0651	0.9941

Таблица 11. Коэффициенты корреляции для пары «Площадь участка» – «Год постройки»

Коэффициенты корреляции Пирсона, Спирмена и тау Кендалла являются значимыми на уровне 5% (0.05), таким образом, мы получаем, что коэффициенты Спирмена и тау Кендалла незначимы, а коэффициенты Пирсона значим. Значение коэффициента говорит о том, что если положительная статистическая связь и существует, то она крайне слабая.

Анализ наличия корреляции между независимыми числовыми и качественными переменными. Чтобы проанализировать пары «числовая независимая переменная – качественная независимая переменная», построим категорированные диаграммы Бокса-Уискера и приведем результаты выполнения теста Краскела-Уолиса.

Анализ пары «Жилая площадь» – «Одноэтажный»

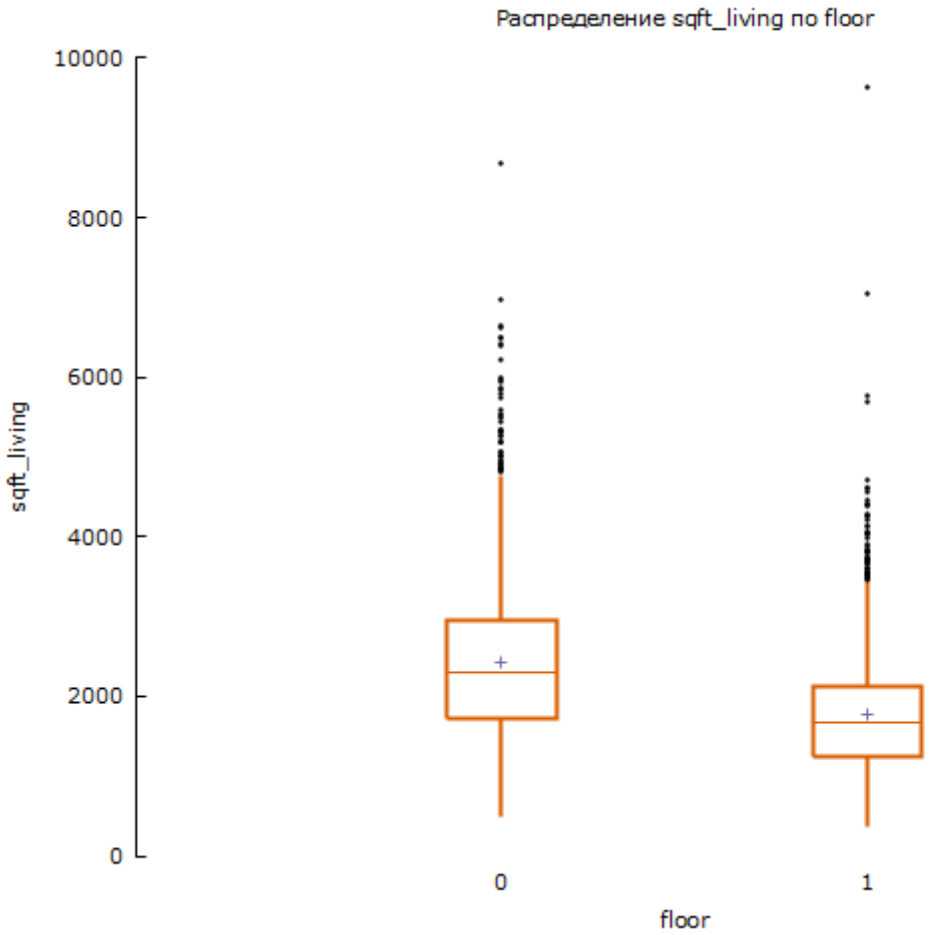

Рисунок 19. Зависимость жилой площади от количества этажей (0 – многоэтажный, 1 – одноэтажный)

Анализируя представленную на рисунке 19 коробчатую диаграмму, можно сделать вывод о том, что медиана значений жилых площадей многоэтажных домов превышает верхний квартиль значений жилых площадей одноэтажных домов. Таким образом, предварительно можно утверждать, что существуют значимые различия в величинах жилой площади между этими группами, и значения жилой площади для первой группы в среднем заметно больше, чем для второй.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 625.4930, df = 1, p-value = 4.77603e-138), который показал,

что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Жилая площадь» – «Подвал»

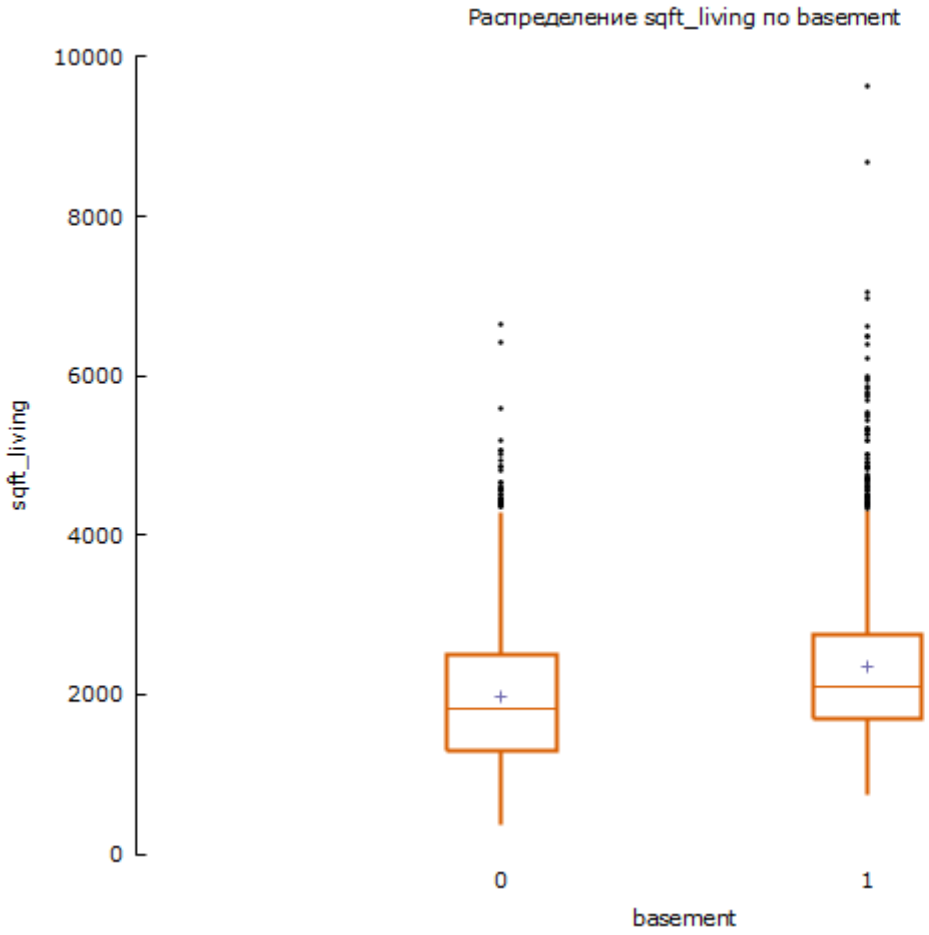

Рисунок 20. Зависимость жилой площади от наличия подвала (0 – нет подвала, 1 – есть подвал)

Анализируя представленную на рисунке 20 коробчатую диаграмму, можно сделать вывод о том, что как медиана, так и разброс значений жилых площадей домов с подвалом несколько превышают медиану и разброс значений жилых площадей домов без подвала. Таким образом, предварительно можно утверждать, что существуют различия в величинах жилой площади между этими группами, и значения жилой площади для второй группы в среднем больше, чем для первой.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 193.6637, df = 1, p-value = 5.04261e-044), который показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Жилая площадь» – «Реконструкция»

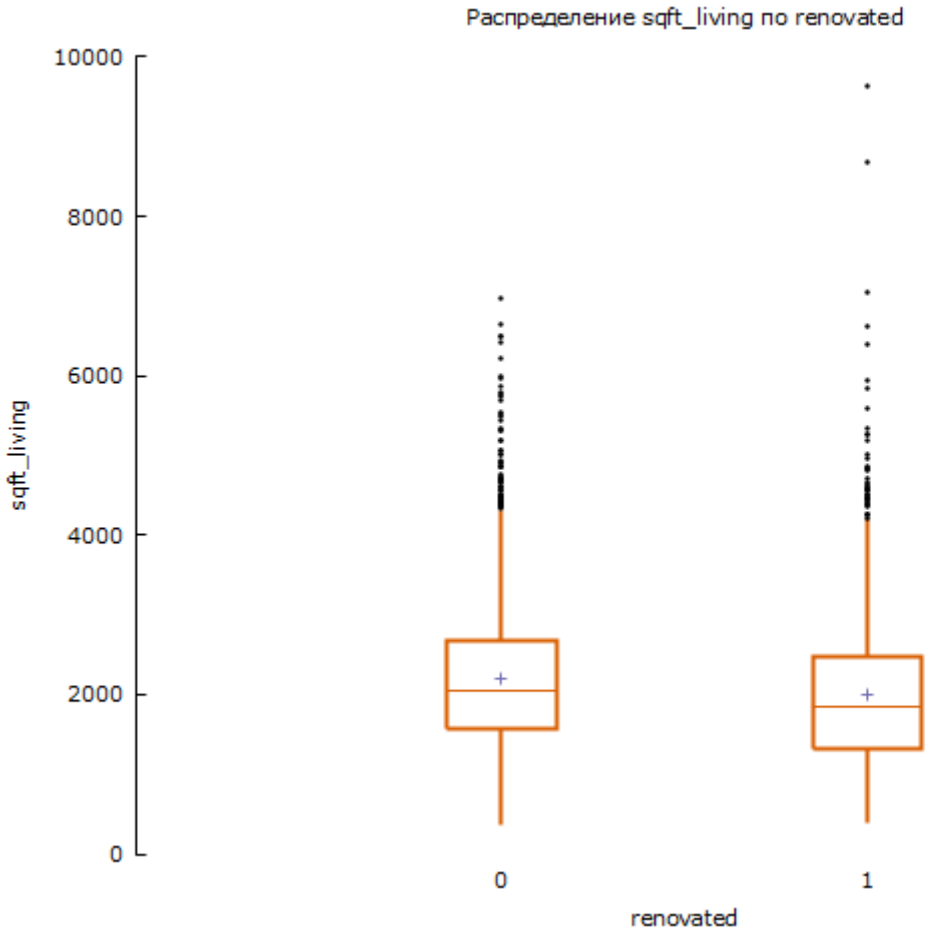

Рисунок 21. Зависимость жилой площади от реконструкции (0 – не было реконструкции, 1 – была реконструкция)

Анализируя представленную на рисунке 21 коробчатую диаграмму, можно сделать вывод о том, что медиана значений жилых площадей отремонтированных домов несколько меньше медианы значений жилых площадей неотремонтированных домов. Таким образом, предварительно можно утверждать, что различия в величинах жилой площади между этими группами могу существовать.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 75.7215, df = 1, p-value = 3.26626e-018), который показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Площадь участка» – «Одноэтажный»

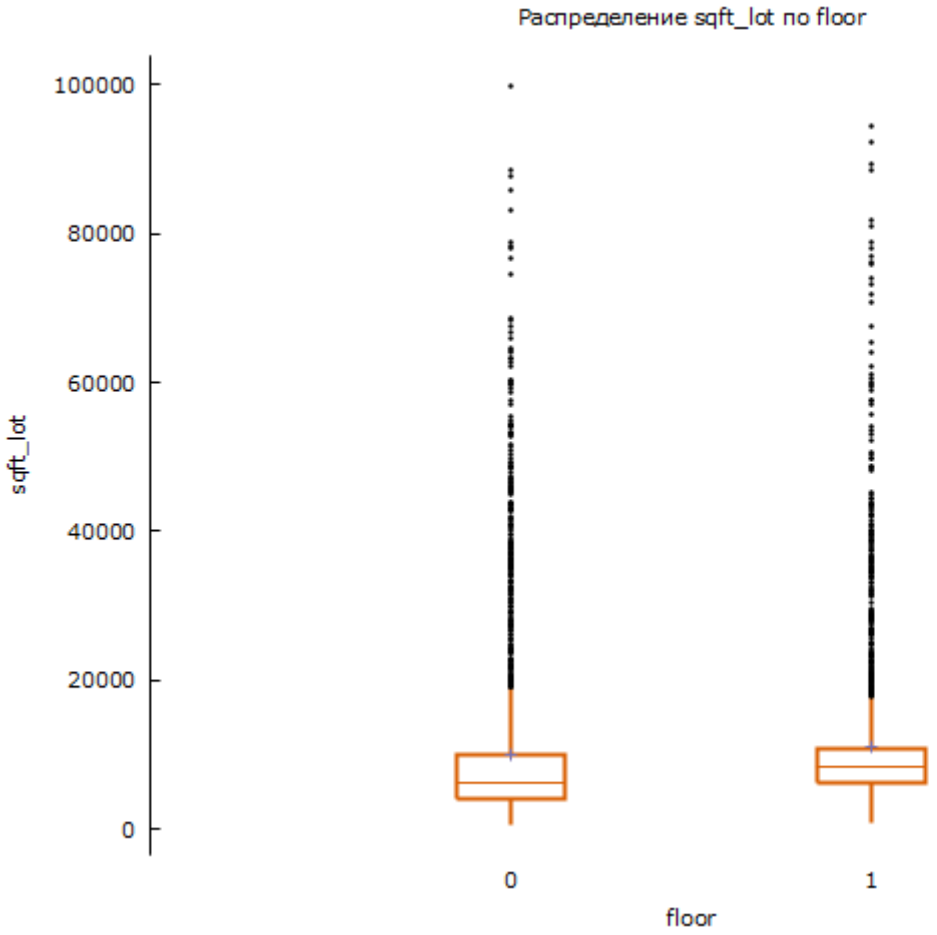

Рисунок 22. Зависимость площади участка от количества этажей (0 – многоэтажный, 1 – одноэтажный)

Анализируя представленную на рисунке 22 коробчатую диаграмму, можно сделать вывод о том, что разброс значений площадей участков многоэтажных домов превышает разброс значений площадей участков одноэтажных домов, но при этом медианы первых меньше медианы вторых. Таким образом, предварительно можно утверждать, что существуют различия в величинах площадей участков между этими группами.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 208.6257, df = 1, p-value = 2.73972e-047), который показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Площадь участка» – «Подвал»

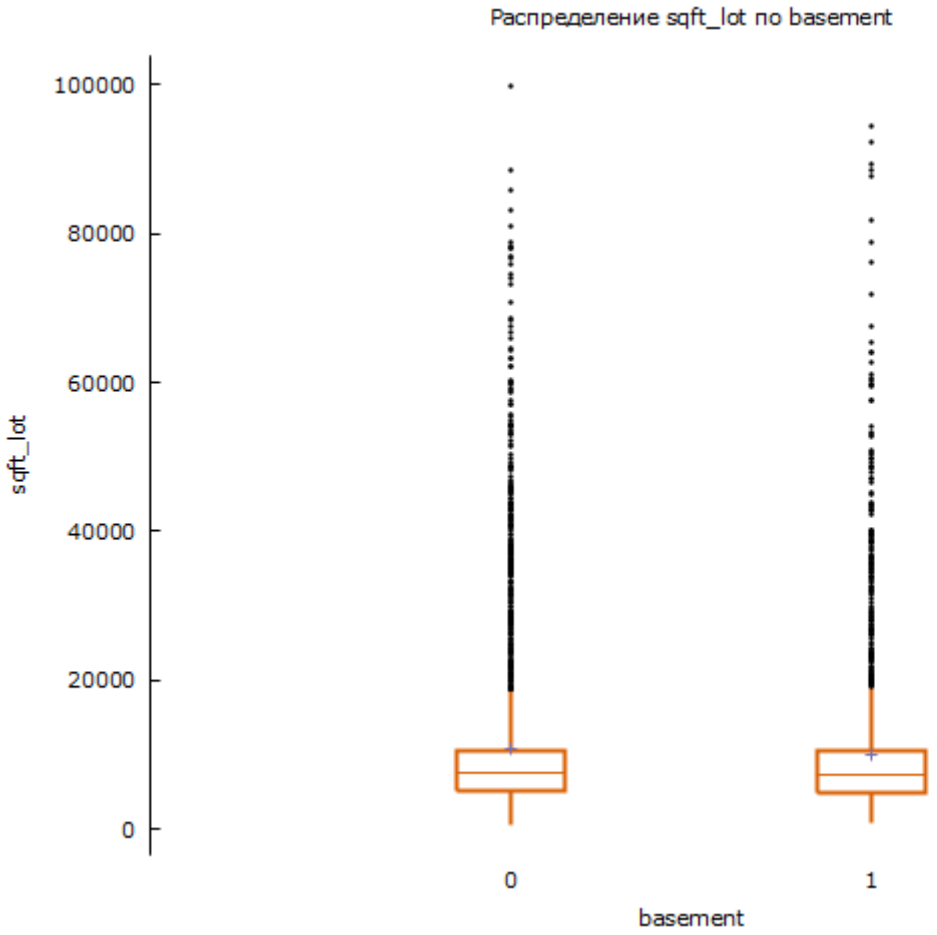

Рисунок 23. Зависимость площади участка от наличия подвала (0 – нет подвала, 1 – есть подвал)

Анализируя представленную на рисунке 23 коробчатую диаграмму, можно сделать вывод о том, что распределение значений площадей участков домов с подвалом схоже с распределением значений площадей участков домов без подвала. Таким образом, предварительно можно утверждать, что если связь между этими переменными и существует, то она крайне слабая.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 5.1586, df = 1, p-value = 0.0231314), который показал, что полученное значение p-value (0.023) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Площадь участка» – «Реконструкция»

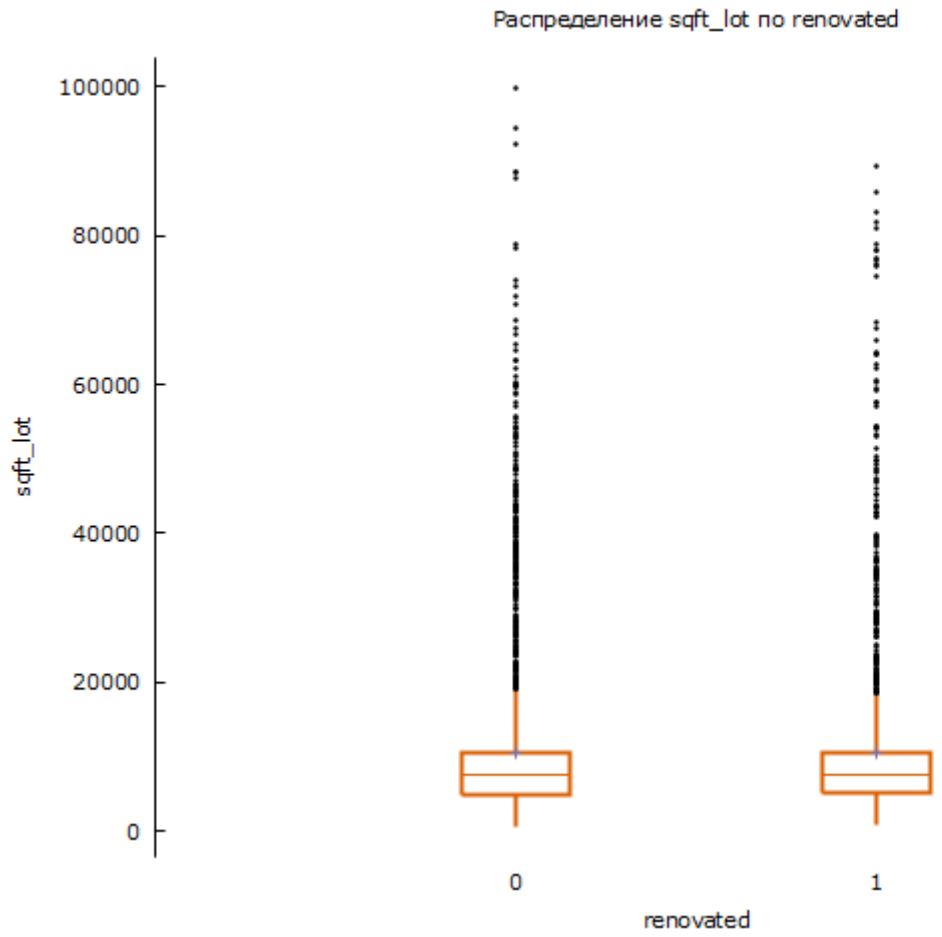

Рисунок 24. Зависимость площади участка от реконструкции (0 – не было реконструкции, 1 – была реконструкция)

Анализируя представленную на рисунке 24 коробчатую диаграмму, можно сделать вывод о том, что распределение значений площадей участков отремонтированных домов схоже с распределением значений площадей участков неотремонтированных домов. Таким образом, предварительно можно утверждать, что если связь между этими переменными и существует, то она крайне слабая.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 7.8996, df = 1, p-value = 0.00494458), который показал, что полученное значение p-value (0.005) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Год постройки» – «Одноэтажный»

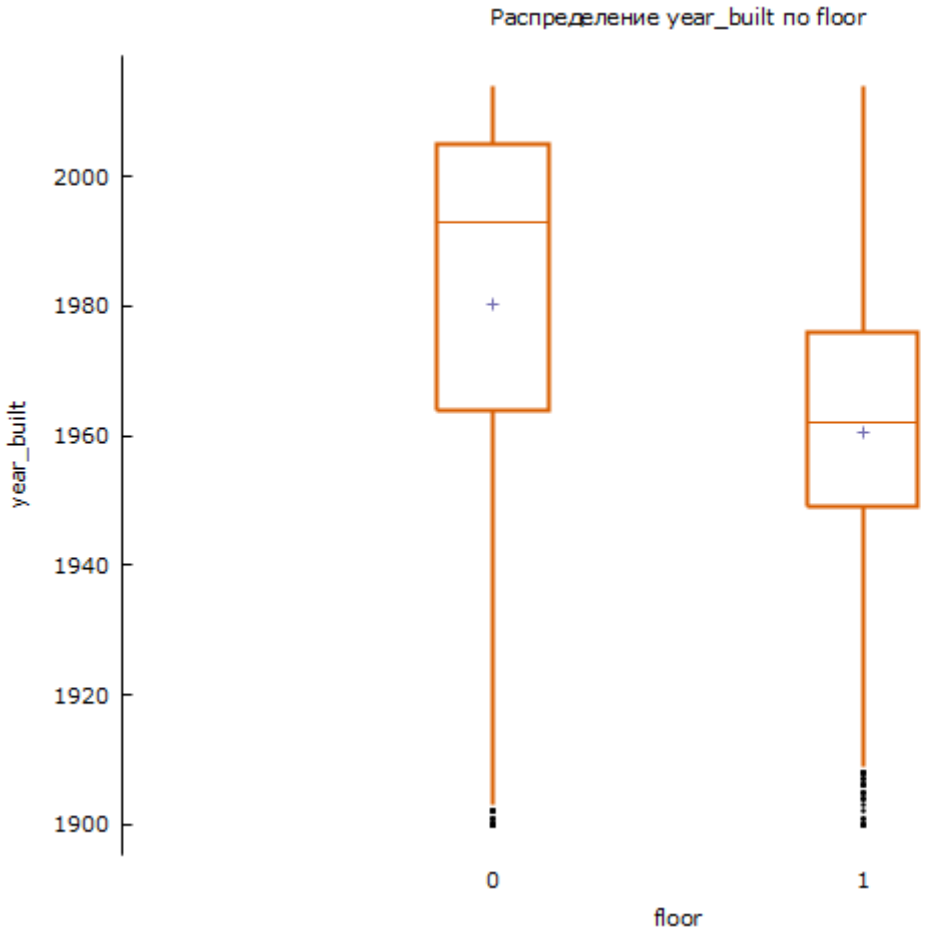

Рисунок 25. Зависимость года постройки от количества этажей (0 – многоэтажный, 1 – одноэтажный)

Анализируя представленную на рисунке 25 коробчатую диаграмму, можно сделать вывод о том, что как медиана, так и разброс значений годов постройки многоэтажных домов превышает медиану и разброс значений годов постройки одноэтажных домов. Таким образом, предварительно можно утверждать, что существуют значимые различия в значениях года постройки между этими группами, и значения года постройки для первой группы в среднем заметно больше, чем для второй (то есть возраст домов первой группы в среднем меньше, чем возраст домов второй группы). В реальности эти результаты можно было бы интерпретировать как то, что раньше чаще строили одноэтажные дома, но в наше время большинство домов строят многоэтажными.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 745.9341, df = 1, p-value = 3.0724e-164), который показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Год постройки» – «Подвал»

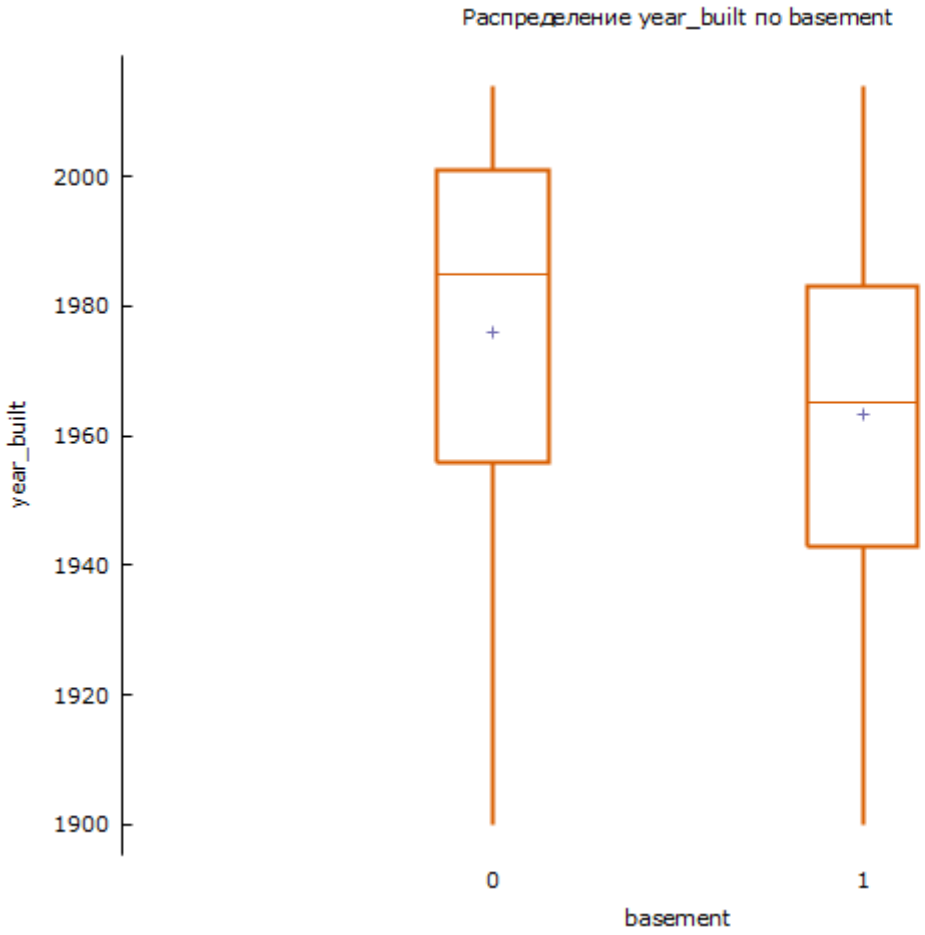

Рисунок 26. Зависимость года постройки от наличия подвала (0 – нет подвала, 1 – есть подвал)

Анализируя представленную на рисунке 26 коробчатую диаграмму, можно сделать вывод о том, что медиана значений годов постройки домов без подвала превышает верхний квартиль значений годов постройки домов с подвалом. Таким образом, предварительно можно утверждать, что существуют значимые различия в значениях года постройки между этими группами, и значения года постройки для первой группы в среднем заметно больше, чем для второй (то есть возраст домов первой группы в среднем меньше, чем возраст домов второй группы). В реальности эти результаты можно было бы интерпретировать как тренд на постройку домов без подвала – мы видим, что среди старых домов чаще встречаются дома с подвалами, а среди новых домов чаще встречаются дома без подвалов.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 211.7005, df = 1, p-value = 5.84604e-048), который показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Анализ пары «Год постройки» – «Реконструкция»

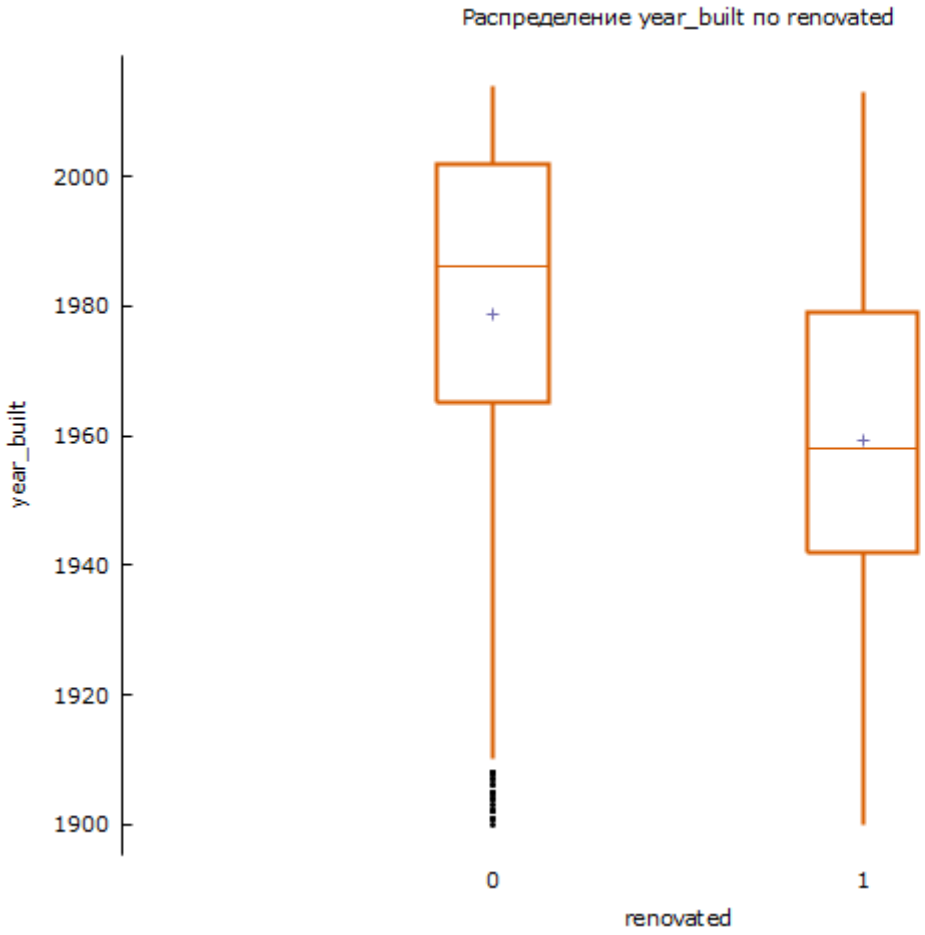

Рисунок 27. Зависимость года постройки от реконструкции (0 – не было реконструкции, 1 – была реконструкция)

Анализируя представленную на рисунке 27 коробчатую диаграмму, можно сделать вывод о том, что медиана значений годов постройки неотремонтированных домов превышает медиану значений годов постройки отремонтированных домов. Таким образом, предварительно можно утверждать, что существуют значимые различия в значениях года постройки между этими группами, и значения года постройки для первой группы в среднем заметно больше, чем для второй (то есть возраст домов первой группы в среднем меньше, чем возраст домов второй группы). Полученный результат вполне очевиден – старые дома ремонтируются чаще, чем новые.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 513.1827, df = 1, p-value = 1.28755e-113), который показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

2.2.4 Предварительная проверка гипотез

Результаты проверки гипотез на основании предварительного анализа данных.

Гипотеза №1: С увеличением возраста дома его цена падает, при этом скорость падения цены отремонтированных домов ниже, чем неотремонтированных.

Из предварительного анализа, проведенного в п. 2.2.2, а именно, из диаграммы рассеивания (рисунок 15) и коэффициентов корреляции Пирсона, Спирмена и тау Кендалла (таблица 5), следует сделать вывод о том, что если связь между возрастом дома и его ценой и существует, то крайне слабая. Дополнительно построим диаграмму рассеивания для этих переменных с разделением факторов (фактор – переменная «Реконструкция»).

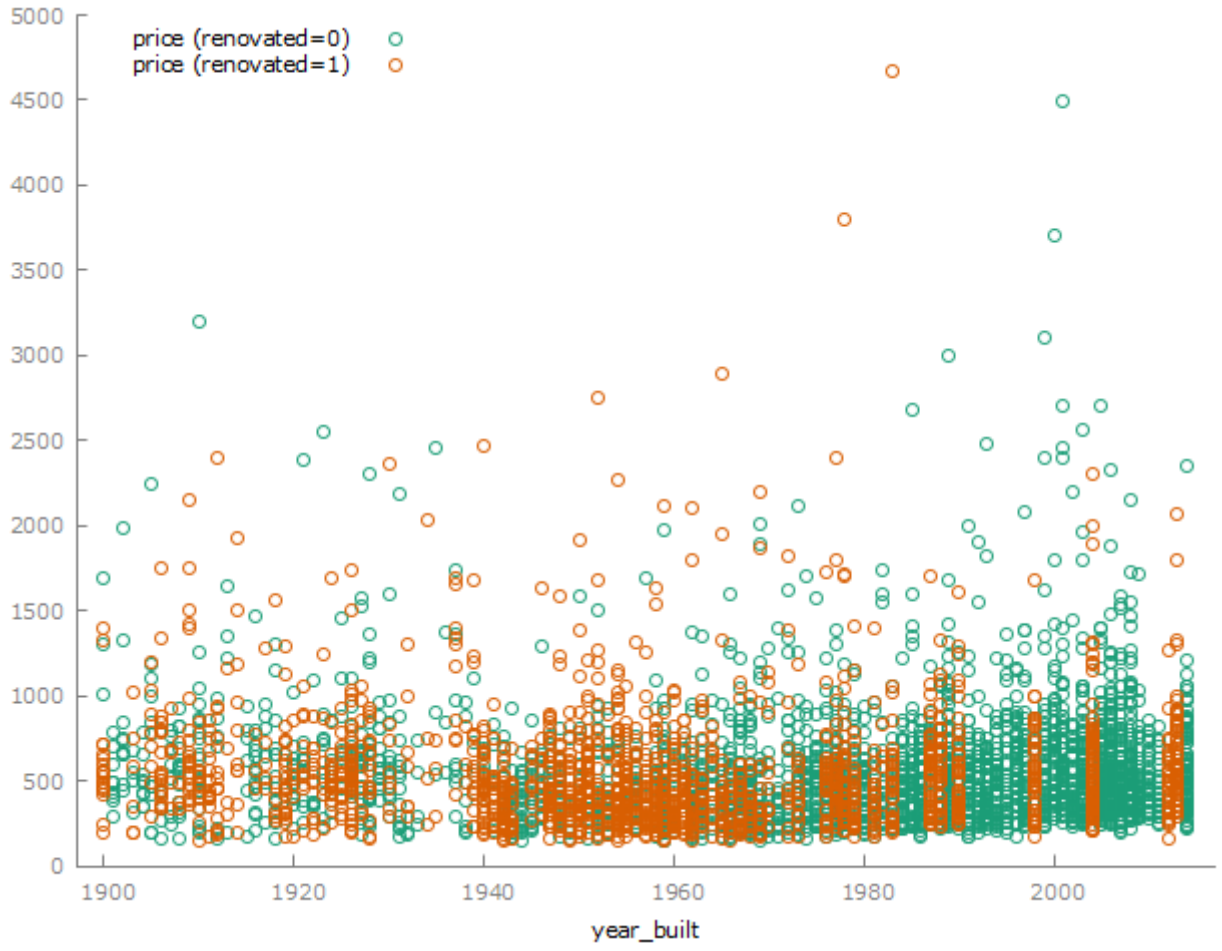

Рисунок 28. Зависимость цены от года постройки с разделением по реконструкции

Анализируя представленный на рисунке 28 график, можно сказать, что между исследуемыми переменными заметная зависимость не усматривается.

Результат проверки: предварительный анализ не подтвердил данную гипотезу.

Гипотеза №2: По мере увеличения площади цена на дом значительно увеличивается до некоторого значения, а потом перестает заметно расти.

Из предварительного анализа, проведенного в п. 2.2.2, а именно, из диаграмм рассеивания (рисунки 12–14) и коэффициентов корреляции Пирсона, Спирмена и тау Кендалла (таблица 4), следует сделать вывод о том, что существует слабая

положительная связь между ценой дома и площадью участка, при этом получено, что цена не растет неограниченно с увеличением площади участка.

Результат проверки: предварительный анализ не опроверг данную гипотезу.

Гипотеза №3: Многоэтажные дома стоят дороже одноэтажных, причем наличие подвала также влияет на рост цены дома, но в меньшей степени.

Для проверки этой гипотезы построим категорированную диаграмму Бокса-Уискера и приведем результаты выполнения теста Краскела-Уолиса.

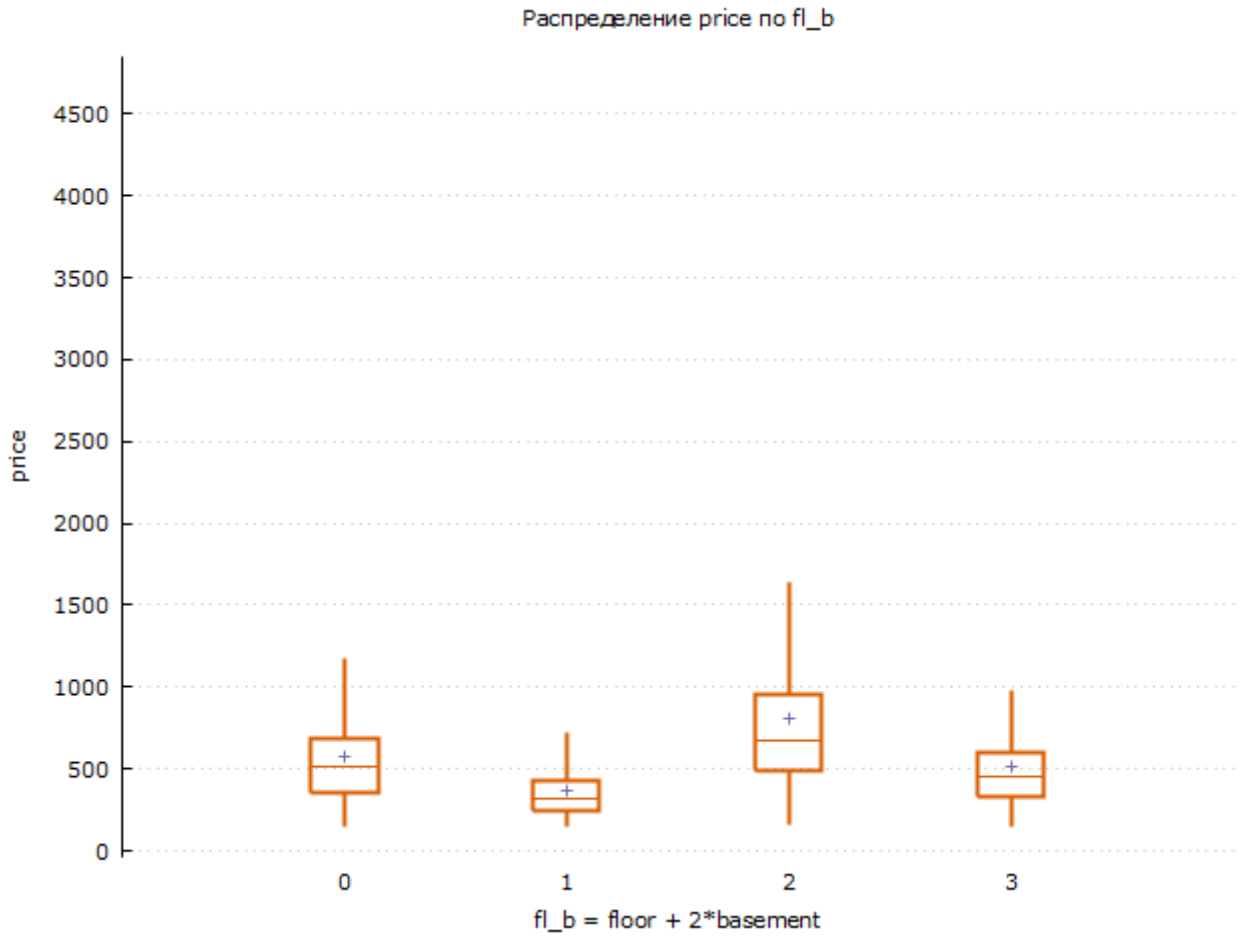

Рисунок 29. Зависимость цены от количества этажей и наличия подвала (0 – многоэтажные дома без подвала, 1 – одноэтажные дома без подвала, 2 – многоэтажные дома с подвалом, 4 – одноэтажные дома с подвалом)

Анализируя представленную на рисунке 29 коробчатую диаграмму, можно сделать вывод о том, что между ценой и количеством этажей, а также наличием подвала существует хорошо заметная связь.

Для формальной проверки гипотезы о наличии статистической связи был выполнен непараметрический дисперсионный анализ с применением теста Краскела-Уоллиса (Kruskal-Wallis chi-squared = 828.0512, df = 3, p-value = 3.56831e-179), который показал, что полученное значение p-value (0.00) меньше допустимого уровня значимости в 5% (0.05), значит нулевая гипотеза об отсутствии связи может быть отвергнута.

Приведем таблицу с числовыми значениями.

fl_b	среднее	Q1	медиана	Q3
0	571.24	359.50	510.00	688.00
1	368.85	245.00	320.75	435.00
2	808.49	487.08	670.50	952.46
3	516.41	336.50	459.00	602.00

Таблица 12. Числовая статистика для price по fl_b

Из таблицы 12 видно, что среднее и медиана значений цен многоэтажных домов независимо от наличия подвала выше, чем среднее и медиана цен одноэтажных. При этом многоэтажные дома с подвалом в среднем стоят дороже многоэтажных домов без подвала, аналогичный тренд можно видеть для одноэтажных домов.

Результат проверки: предварительный анализ не опроверг данную гипотезу.

3 Спецификация, оценивание и оптимизация модели

3.1 Спецификация моделей для проверки гипотез и решения поставленной задачи

В качестве базовой модели используется регрессионная модель с линейным включением всех переменных.

Модель имеет следующий вид:

$$price = a_0 + a_1 * sqft\_living + a_2 * sqft\_lot + a_3 * floor + a_4 * basement + a_5 * year\_built + a_6 * renovated.$$

Гипотезу 3 о том, что увеличение количества этажей и наличие подвала влияет на цену в сторону ее увеличения, проверим следующим образом. Если гипотеза о том, что количество этажей и наличие подвала не влияют на цену, верна, то коэффициенты при переменных floor и basement $a_3 = 0$ и $a_4 = 0$ соответсвенно. Если же верна альтернативная гипотеза о том, что чем больше этажей, тем дороже дом, и если есть подвал, то цена выше, то $a_3 < 0$ и $a_4 > 0$. При этом из предположения о том, что наличие подвала, как и увеличение количества этажей, влияет на рост цены, но в меньшей степени, следует, что $|a_3| > |a_4|$.

Гипотезу 1 о том, что увеличение года постройки дома связано с увеличением цены на него, проверим следующим образом. Если гипотеза о том, что год постройки не влияет на цену, верна, то коэффициент при переменной year_built $a_5 = 0$. Если же верна альтернативная гипотеза о том, что чем позже построен дом, тем выше его цена, то $a_5 > 0$.

Вторую часть гипотезы 1 о том, что если дом был на реконструкции, то его цена падает медленнее с увеличением возраста дома, проверим следующим образом. Для этого учтем, что коэффициент перед годом постройки зависит от реконструкции следующим образом: $a_5(renovated) = b_1 * renovated + b_2$. Запишем, что $a_5(renovated) > 0$, поскольку чем позже построен дом, тем выше его цена. Тогда если $renovated = 0$, то $a_5 = b_2 > 0$, иначе при $renovated = 1$ $a_5 = b_1 + b_2 > 0$. Так как в гипотезе предполагается, что отремонтированные дома дешевеют медленнее, то $b_1 > 0$.

Тогда запишем выражение для регрессии, модифицировав базовую модель:

$$price = a_0 + a_1 * sqft\_living + a_2 * sqft\_lot + a_3 * floor + a_4 * basement + b_1 * renovated * year\_built + b_2 * year\_built.$$

Гипотезу 2 о том, что увеличение общей площади связано с увеличением цены на него, проверим следующим образом. Если гипотеза о том, что общая площадь не влияет на цену, верна, то коэффициент при переменной sqft_lot $a_2 = 0$. Если же верна альтернативная гипотеза о том, что рост общей площади влияет на цену в сторону ее увеличения, то $a_2 > 0$.

Вторую часть гипотезы 2 о том, что начиная с какого-то значения площади участка цена перестает расти, проверим следующим образом. Для этого проведем модификацию модели. Зададим переменную-индикатор $sqft\_lot\_ind = 1 * (sqft\_lot < 20000) + 0 * (sqft\_lot \geq 20000)$. Примерное пороговое значение $sqft\_lot = 20000$ обусловлено результатами, полученными в пункте 2.2.2 (рисунки 13, 14). Тогда значение коэффициента при переменной sqft_lot зависит от переменной-индикатора sqft_lot_ind, то есть $a_2(sqft\_lot\_ind) = c_1 * sqft\_lot\_ind + c_2$. Запишем, что $a_2(sqft\_lot\_ind) > 0$, поскольку чем больше площадь участка, тем выше цена дома. Тогда если $sqft\_lot\_ind = 0$, то $a_2 = c_2 > 0$, иначе при $sqft\_lot\_ind = 1$ $a_2 = c_1 + c_2 > 0$. Так как в гипотезе предполагается, что до некоторого значения площади участка цена заметно растет, то $c_1 > 0$.

Тогда запишем выражение для регрессии, модифицировав модель для гипотезы 1:

$$price = a_0 + a_1 * sqft\_living + c_1 * sqft\_lot\_ind * sqft\_lot + c_2 * sqft\_lot + a_3 * floor + a_4 * basement + b_1 * renovated * year\_built + b_2 * year\_built.$$

3.2 Оценивание базовой модели и результаты проверки гипотез

В данном разделе выборка была случайным образом разделена на две части: обучающую (80% от общего объема данных) и тестовую (20%). Обучающая часть будет использована для построения модели, тестовая – для проверки прогностических свойств.

Анализ базовой модели

Базовая модель выглядит следующим образом:

$$price = a_0 + a_1 * sqft\_living + a_2 * sqft\_lot + a_3 * floor + a_4 * basement + a_5 * year\_built + a_6 * renovated.$$

Базовая модель: МНК, использованы наблюдения 1–3520.

Зависимая переменная: price.

Переменная	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	5044,99	307,994	16,38	<0,0001
sqft_living	0,290100	0,00566665	51,19	<0,0001
sqft_lot	−0,00182137	0,000395277	−4,608	<0,0001
floor	−40,2316	9,95415	−4,042	<0,0001
basement	−5,12807	9,58433	−0,5350	0,5927
year_built	−2,57002	0,156237	−16,45	<0,0001
renovated	−10,0090	8,86478	−1,129	0,2589

Среднее завис. перемен	547,0364	Ст. откл. завис. перем	350,8381
Сумма кв. остатков	2,08e+08	Ст. ошибка модели	243,3612
R-квадрат	0,519661	Исправ. R-квадрат	0,518840
F(6, 4393)	633,4302	P-значение (F)	0,000000
Лог. правдоподобие	−24331,96	Крит. Акаике	48677,93
Крит. Шварца	48721,09	Крит. Хеннана-Куинна	48693,33

Таблица 13. Базовая модель

Интерпретация полученных коэффициентов:

- Const = 5044.99.
- Коэффициент перед sqft_living равен 0.290100. Полученный результат можно интерпретировать как то, что при прочих равных в среднем дома с большей жилой площадью стоят дороже.
- Коэффициент перед sqft_lot равен −0.00182137. Полученный результат можно интерпретировать как то, что при прочих равных в среднем дома с большей площадью участка стоят дешевле.
- Коэффициент перед floor равен −40.2316. Полученный результат можно интерпретировать как то, что при прочих равных в среднем одноэтажные дома стоят на 40.2 тыс. долларов дешевле многоэтажных.
- Коэффициент перед basement равен −5.12807. Полученный результат можно интерпретировать как то, что при прочих равных в среднем дома с подвалом стоят на 5.1 тыс. долларов дешевле домов без подвала.
- Коэффициент перед year_built равен −2.57002. Полученный результат можно интерпретировать как то, что при прочих равных в среднем дома, построенные позже, стоят меньше.
- Коэффициент перед renovated равен −10.0090. Полученный результат можно интерпретировать как то, что при прочих равных в среднем реконструированные дома стоят на 10 тыс. долларов дешевле нереконструированных.

Как и предполагалось, увеличение жилой площади дома и количества этажей влияет на рост цены, а факт того, что дом был на реконструкции, уменьшает цену. Коэффициент перед переменой basement отрицательный, значит, наличие подвала уменьшает цену, что противоречит предположениям и ранее полученным результатам. Однако площадь подвала входит в жилую площадь дома, поэтому подвал площадью более чем 18 кв. футов (0.290100 * 18–5.12807 * 1 = 0.094>0) увеличивает стоимость. То есть не столько наличие подвала влияет на цену, сколько тот факт, что его площадь учитывается в жилой площади дома. Значение const и коэффициент перед переменной year_built кажутся противоречивыми. Однако эти результаты можно объяснить особенностью выборки, а именно диапазоном годов постройки домов (1900–2014). Таким образом, большая часть константы нивелируется значением года постройки. Коэффициент перед

переменной sqft_lot отрицательный, но достаточно малый, что говорит о медленном снижении стоимости дома при увеличении метража участка.

Заметим, что все коэффициенты являются значимыми на уровне 5% (0.05), таким образом, мы получаем, что коэффициенты перед переменными basement и renovated незначимы. В связи с тем, что выше полученные результаты указывают на то, что важен не сам факт наличия подвала, а его площадь, которая учитывается в жилой площади дома, было принято решение удалить переменную basement из модели.

Переходим к проверке гипотезы 3. Как и предполагалось, коэффициент перед переменной floor $a_3 < 0$, коэффициент a_4 перед переменной basement также отрицательный (но он незначим), а неравенство $|a_3| > |a_4|$ выполнено, из чего можно сделать вывод о том, что гипотеза 3 верна.

Анализ обновленной базовой модели

Обновленная базовая модель без переменной basement выглядит следующим образом:

$price = a_0 + a_1 * sqft\_living + a_2 * sqft\_lot + a_3 * floor + a_5 * year\_built + a_6 * renovated.$

Обновленная базовая модель: МНК, использованы наблюдения 1–3520.

Зависимая переменная: price.

Переменная	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	5010,98	301,331	16,63	<0,0001
sqft_living	0,288908	0,00520960	55,46	<0,0001
sqft_lot	−0,00178316	0,000388732	−4,587	<0,0001
floor	−42,1166	9,30888	−4,524	<0,0001
year_built	−2,55232	0,152678	−16,72	<0,0001
renovated	−9,86458	8,85978	−1,113	0,2656

Среднее завис. перемен	547,0364	Ст. откл. завис. перем	350,8381
Сумма кв. остатков	2,08e+08	Ст. ошибка модели	243,3365
R-квадрат	0,519622	Исправ. R-квадрат	0,518938
F(5, 3514)	760,2134	P-значение (F)	0,000000
Лог. правдоподобие	−24332,11	Крит. Акаике	48676,22
Крит. Шварца	48713,21	Крит. Хеннана-Куинна	48689,42

Таблица 14. Обновленная базовая модель

В обновленной модели несколько увеличилось значение adjR^2 и несколько уменьшилось значение критерия Акаике.

Для проверки модели на мультиколлинеарность подсчитаем значения коэффициента VIF для всех регрессоров и проведем диагностику коллинеарности Белсли-Ку-Велша. Значения коэффициента VIF приведены в таблице 15.

Переменная	Коэффициент VIF
sqft_living	1,304
sqft_lot	1,124
floor	1,283
year_built	1,234
renovated	1,122

Таблица 15. Значения коэффициента VIF для регрессоров обновленной базовой модели.

Все значения коэффициента VIF небольшие, близкие к единице, что говорит об отсутствии мультиколлинеарности.

Результаты диагностики коллинеарности Белсли-Ку-Велша приведены в таблице 16.

lambda	cond	const	sqft_living	sqft_lot	floor	year_built	renovated
4,430	1,000	0,000	0,005	0,015	0,012	0,000	0,014
0,641	2,630	0,000	0,021	0,149	0,163	0,000	0,344
0,452	3,131	0,000	0,011	0,031	0,521	0,000	0,415
0,402	3,322	0,000	0,018	0,735	0,000	0,000	0,151
0,075	7,672	0,000	0,930	0,067	0,249	0,000	0,008
0,000	218,509	1,000	0,015	0,003	0,055	1,000	0,067

Таблица 16. Диагностика коллинеарности Белсли-Ку-Велша для обновленной базовой модели

Анализируя значения индексов обусловленности, можно прийти к выводу, что существует умеренная зависимость между переменными sqft_living и floor (подобные результаты уже были получены в пункте 2.2.3), а также сильная зависимость между const и year_built (что также было отмечено при анализе базовой модели). В связи с тем, что исключение регрессора year_built ухудшило характеристики модели, было принято решение не удалять его из модели.

При дальнейшем модифицировании модели было принято решение исключить переменную renovated в связи с ее незначимостью, что улучшило характеристики модели.

Анализ модифицированной модели для проверки гипотезы 1

Модифицированная модель для проверки гипотезы 1 выглядит следующим образом:

$price = a_0 + a_1 * sqft\_living + a_2 * sqft\_lot + a_3 * floor + b_1 * renovated * year\_built + b_2 * year\_built.$

Модифицированная модель 1: МНК, использованы наблюдения 1–3520.

Зависимая переменная: price.

Здесь $year\_ren = renovated * year\_built.$

Переменная	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	5009,42	300,421	16,67	<0,0001
sqft_living	0,288912	0,00520956	55,46	<0,0001
sqft_lot	−0,00178238	0,000388739	−4,585	<0,0001
floor	−42,0717	9,30992	−4,519	<0,0001
year_ren	−0,00515438	0,00450449	−1,144	0,2526
year_built	−2,55150	0,152236	−16,76	<0,0001

Среднее завис. перемен	547,0364	Ст. откл. завис. перем	350,8381
Сумма кв. остатков	2,08e+08	Ст. ошибка модели	243,3341
R-квадрат	0,519631	Исправ. R-квадрат	0,518948
F(5, 3514)	760,2424	P-значение (F)	0,000000
Лог. правдоподобие	−24332,07	Крит. Акаике	48676,15
Крит. Шварца	48713,14	Крит. Хеннана-Куинна	48689,35

Таблица 17. Модифицированная модель 1

В модифицированной модели 1 несколько увеличилось значение adjR^2 и несколько уменьшилось значение критерия Акаике.

Заметим, что все коэффициенты являются значимыми на уровне 5% (0.05), таким образом, мы получаем, что коэффициент перед переменной year_ren незначим. В связи с тем, что исключение этого регрессора не приводило к улучшению характеристик, было принято решение не удалять переменную year_ren из модели.

Переходим к проверке гипотезы 1. Коэффициент перед переменной year_built $b_2 < 0$, а коэффициент b_2 при переменной year_ren также отрицательный, что противоречит предположениям гипотезы 1, из чего можно сделать вывод о том, что гипотеза 1 неверна.

Для проверки модели на мультиколлинеарность подсчитаем значения коэффициента VIF для всех регрессоров и проведем диагностику коллинеарности Белсли-Ку-Велша. Значения коэффициента VIF приведены в таблице 18.

Переменная	Коэффициент VIF
sqft_living	1,304
sqft_lot	1,125

floor	1,284
year_ren	1,114
year_built	1,227

Таблица 18. Значения коэффициента VIF для регрессоров модифицированной модели 1

Все значения коэффициента VIF небольшие, близкие к единице, что говорит об отсутствии мультиколлинеарности.

Результаты диагностики коллинеарности Белсли-Ку-Велша приведены в таблице 19.

lambda	cond	const	sqft_living	sqft_lot	floor	year_ren	year_built
4,431	1,000	0,000	0,005	0,015	0,012	0,015	0,000
0,639	2,634	0,000	0,021	0,149	0,164	0,347	0,000
0,453	3,127	0,000	0,011	0,027	0,519	0,425	0,000
0,402	3,320	0,000	0,019	0,739	0,000	0,144	0,000
0,075	7,671	0,000	0,930	0,067	0,250	0,008	0,000
0,000	217,971	1,000	0,015	0,003	0,055	0,062	1,000

Таблица 19. Диагностика коллинеарности Белсли-Ку-Велша для модифицированной модели 1

Анализируя значения индексов обусловленности, можно прийти к выводу, что существует умеренная зависимость между переменными sqft_living и floor (подобные результаты уже были получены в пункте 2.2.3), а также сильная зависимость между const и year_built (что также было отмечено при анализе базовой модели). В связи с тем, что исключение регрессора year_built ухудшило характеристики модели, было принято решение не удалять его из модели.

Анализ модифицированной модели для проверки гипотезы 2

Модифицированная модель для проверки гипотезы 2 выглядит следующим образом:

$price = a_0 + a_1 * sqft\_living + c_1 * sqft\_lot\_ind * sqft\_lot + c_2 * sqft\_lot + a_3 * floor + b_1 * renovated * year\_built + b_2 * year\_built.$

Модифицированная модель 2: МНК, использованы наблюдения 1–3520.

Зависимая переменная: price.

Здесь $year\_ren = renovated * year\_built$, $sl\_sl\_ind = sqft\_lot\_ind * sqft\_lot.$

Переменная	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	5005,76	299,532	16,71	<0,0001
sqft_living	0,295197	0,00536478	55,03	<0,0001
sl_sl_ind	−0,00491793	0,00105032	−4,682	<0,0001
sqft_lot	−0,00226988	0,000401327	−5,656	<0,0001

floor	−29,3552	9,67148	−3,035	0,0024
year_ren	−0,00464971	0,00449243	−1,035	0,3007
year_built	−2,54046	0,151803	−16,74	<0,0001

Среднее завис. перемен	547,0364	Ст. откл. завис. перем	350,8381
Сумма кв. остатков	2,07e+08	Ст. ошибка модели	242,6128
R-квадрат	0,522610	Исправ. R-квадрат	0,521795
F(6, 3513)	640,9617	P-значение (F)	0,000000
Лог. правдоподобие	−24321,12	Крит. Акаике	48656,25
Крит. Шварца	48699,41	Крит. Хеннана-Куинна	48671,65

Таблица 20. Модифицированная модель 2

В модифицированной модели 2 увеличилось значение adjR^2 и уменьшилось значение критерия Акаике.

Заметим, что все коэффициенты являются значимыми на уровне 5% (0.05), таким образом, мы получаем, что коэффициент перед переменной year_ren незначим. В связи с тем, что исключение этого регрессора не приводило к улучшению характеристик, было принято решение не удалять переменную year_ren из модели.

Переходим к проверке гипотезы 2. Коэффициент перед переменной sl_sl_ind $c_1 < 0$, а коэффициент c_2 при переменной sqft_lot также отрицательный, что противоречит предположениям гипотезы 2. Эта гипотеза утверждала, что цена дома значительно растет с увеличением площади участка до некоторого значения, а после перестает заметно расти. Полученные коэффициенты модели указывают на обратную зависимость цены от метража участка, но при этом подтверждают существование точки, знаменующей начало уменьшения силы влияния регрессора на зависимую переменную. Исходя из полученных результатов, нельзя сказать, что гипотеза 2 верна.

Для проверки модели на мультиколлинеарность подсчитаем значения коэффициента VIF для всех регрессоров и проведем диагностику коллинеарности Белсли-Ку-Велша. Значения коэффициента VIF приведены в таблице 21.

Переменная	Коэффициент VIF
sqft_living	1,391
sl_sl_ind	1,152
sqft_lot	1,206
floor	1,394
year_ren	1,115
year_built	1,227

Таблица 21. Значения коэффициента VIF для регрессоров модифицированной модели 2

Все значения коэффициента VIF небольшие, близкие к единице, что говорит об отсутствии мультиколлинеарности.

Результаты диагностики коллинеарности Белсли-Ку-Велша приведены в таблице 22.

lambda	cond	const	sqft_living	sl_sl_ind	sqft_lot	floor	year_ren	year_built
5,163	1,000	0,000	0,004	0,007	0,009	0,008	0,010	0,000
0,640	2,840	0,000	0,020	0,002	0,168	0,146	0,312	0,000
0,496	3,227	0,000	0,003	0,118	0,399	0,001	0,273	0,000
0,452	3,379	0,000	0,014	0,002	0,090	0,475	0,320	0,000
0,174	5,450	0,000	0,020	0,849	0,242	0,050	0,016	0,000
0,074	8,335	0,000	0,926	0,022	0,090	0,268	0,007	0,000
0,000	235,203	0,999	0,014	0,000	0,003	0,052	0,062	0,999

Таблица 22. Диагностика коллинеарности Белсли-Ку-Велша для модифицированной модели 2

Анализируя значения индексов обусловленности, можно прийти к выводу, что существует умеренная зависимость между переменными sqft_living и floor (подобные результаты уже были получены в пункте 2.2.3), а также сильная зависимость между const и year_built (что также было отмечено при анализе базовой модели). В связи с тем, что исключение регрессора year_built ухудшило характеристики модели, было принято решение не удалять его из модели.

3.3 Анализ наличия выбросов

В соответствии с графиком Леверижда на рисунок 30 видно, что все объекты в равной степени влияют на уравнение регрессии.

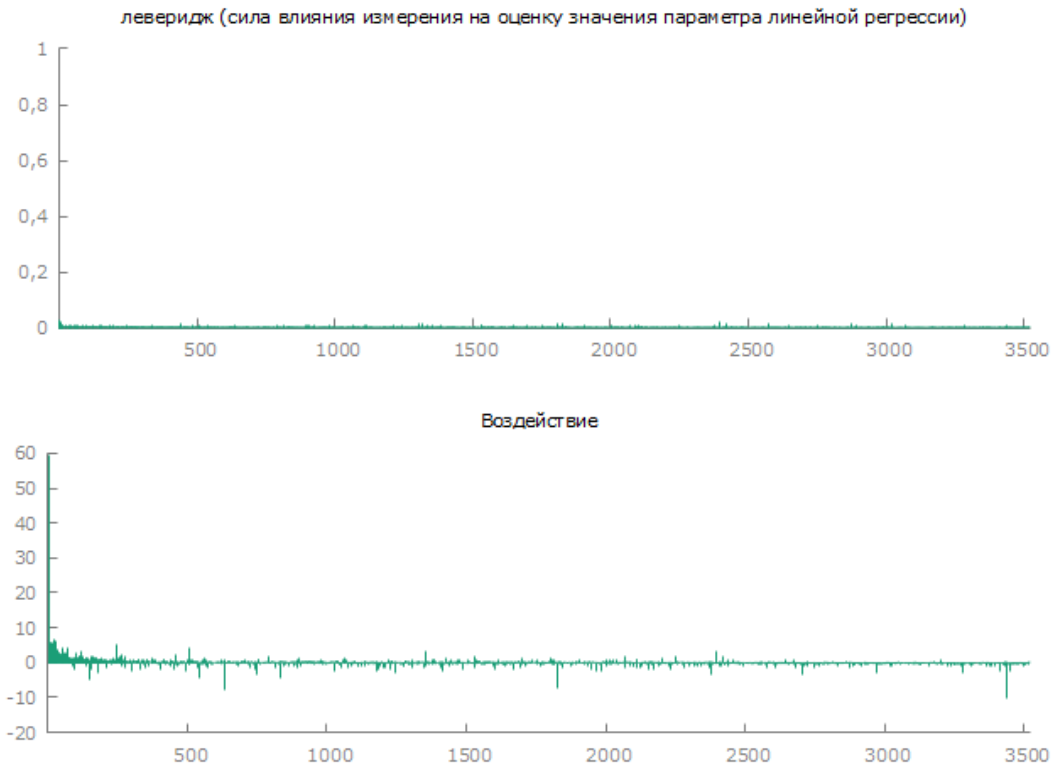

Рисунок 30. Графики Левериджа и Воздействия для модифицированной модели 2

Выделим в отдельную таблицу измерения (выбросы), соответствующие точкам разбалансировки, которые выделяются на фоне других значений Леверижда. О потенциальных выбросах можно говорить и при анализе статистики Воздействия. Измерения с большим значением статистики Воздействия также занесем в таблицу и дополнительно проанализируем.

Так как выборка достаточно большая, то выбросов получилось довольно много (порядка 6% от объема обучающей части). В связи с этим в таблице 23 приведем некоторые из них.

№	sqft_living	sqft_lot	floor	basement	year_built	renovated	price
1	9640	13068	1	1	1983	1	4668,000
2	6430	27517	0	0	2001	0	4489,000
3	7050	42840	1	1	1978	1	3800,000
4	5550	28078	0	1	2000	0	3710,000
5	6210	8856	0	1	1910	0	3200,000
8	8670	64033	0	1	1965	1	2888,000
10	5305	8401	0	1	2005	0	2700,000

Таблица 23. Выбросы

Проанализировав наблюдения, отмеченные как выбросы, можно сказать, что, например, данные, приведенные в таблице 23, были отброшены, как измерения с очень высокой ценой (более чем 2500 тыс. долларов). Также среди выбросов есть много домов с крайне большой жилой площадью. Также к выбросам были отнесены те категории домов, которые при огромной общей площади стоят достаточно мало.

Удалим наблюдения, отмеченные как выбросы, из выборки и построим модифицированную модель гипотезы 2 еще раз.

Модифицированная модель 2: МНК, использованы наблюдения 1–3312.

Зависимая переменная: price.

Здесь $year\_ren = renovated * year\_built$, $sl\_sl\_ind = sqft\_lot\_ind * sqft\_lot$.

Переменная	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	4756,34	274,644	17,32	<0,0001
sqft_living	0,265295	0,00585259	45,33	<0,0001
sl_sl_ind	−0,00417086	0,00105689	−3,946	<0,0001
sqft_lot	−0,00240879	0,000565877	−4,257	<0,0001
floor	−41,4248	9,23874	−4,484	<0,0001
year_ren	−0,00578549	0,00413570	−1,399	0,1619
year_built	−2,38381	0,139397	−17,10	<0,0001

Среднее завис. перемен	523,3349	Ст. откл. завис. перем	292,5799

Сумма кв. остатков	1,53e+08	Ст. ошибка модели	215,2581
R-квадрат	0,459691	Исправ. R-квадрат	0,458710
F(6, 3513)	468,6455	P-значение (F)	0,000000
Лог. правдоподобие	−22487,55	Крит. Акаике	44989,09
Крит. Шварца	45031,83	Крит. Хеннана-Куинна	45004,39

Таблица 24. Модифицированная модель 2 без выбросов

Из таблицы 24 видно, что значимость всех коэффициентов возросла, а стандартные ошибки уменьшились. Критерий Акаике уменьшился, что говорит о том, что модель лучше описывает данные, однако критерий adjR^2 также уменьшился.

3.4 Анализ наличия гетероскедастичности

В таблице 25 представлены результаты проверки модифицированной модели 2 на гетероскедастичность с помощью критерия Уайта (нулевая гипотеза – гомоскедастичность, альтернативная гипотеза – гетероскедастичность).

Тестовая статистика	TR^2 = 915,334306
P-значение	P(Хи-квадрат(26) > 915,334306) = 0,000000

Таблица 25. Тест Уайта на гетероскедастичность

Анализируя результат теста Уайта, представленный в таблице 26, можно сделать вывод о том, что нулевая гипотеза об отсутствии гетероскедастичности может быть отвергнута. Таким образом, гетероскедастичность есть, то есть неверно предположение о том, что дисперсии случайных ошибок совпадают. Воспользуемся корректировкой Уайта, для этого пересчитаем модифицированную модель 2, используя робастные стандартные ошибки.

Модифицированная модель 2: МНК, использованы наблюдения 1–3520.

Зависимая переменная: price.

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC0

Здесь $year\_ren = renovated * year\_built$, $sl\_sl\_ind = sqft\_lot\_ind * sqft\_lot$.

Переменная	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	5005,76	329,825	15,18	<0,0001
sqft_living	0,295197	0,0133449	22,12	<0,0001
sl_sl_ind	−0,00491793	0,00143860	−3,419	0,0006
sqft_lot	−0,00226988	0,000403829	−5,621	<0,0001
floor	−29,3552	12,2191	−2,402	0,0163
year_ren	−0,00464971	0,00431713	−1,077	0,2815
year_built	−2,54046	0,169824	−14,96	<0,0001

Среднее завис. перемен	547,0364	Ст. откл. завис. перем	350,8381
Сумма кв. остатков	2,07e+08	Ст. ошибка модели	242,6128
R-квадрат	0,522610	Исправ. R-квадрат	0,521795
F(6, 3513)	157,5326	P-значение (F)	0,000000
Лог. правдоподобие	−24321,12	Крит. Акаике	48656,25
Крит. Шварца	48699,41	Крит. Хеннана-Куинна	48671,65

Таблица 26. Модифицированная модель 2 с поправкой на гетероскедастичность

Из таблицы 26 видно, что несколько уменьшилась значимость некоторых коэффициентов, при этом заметно увеличились значения стандартных ошибок, то есть все оценки стали менее точными.

3.5 Оптимизация модели

В таблица 27 представлены основные характеристики, определяющие оптимальность построенных моделей (за исключением модифицированной модели 2 без выбросов), а именно – коэффициент детерминации R-квадрат, (и исправ. R-квадрат) и критерий Акаике.

Модель/Критерий	R^2	adjR^2	Akaike
Базовая модель	0,519661	0,518840	48677,93
Обновленная базовая модель	0,519622	0,518938	48676,22
Модифицированная модель 1	0,519631	0,518948	48676,15
Модифицированная модель 2	0,522610	0,521795	48656,25

Таблица 27. Результаты анализа оптимальности построенных моделей

Анализируя значения критериев, представленных в таблице 27, можно сделать вывод о том, что наилучшей моделью с точки зрения наибольшего значения коэффициента детерминации и наименьшего значения коэффициента Акаике является модифицированная модель 2.

3.6 Проверка прогностических свойств модели

На рисунке 31 представлены прогнозы значений зависимой переменной для тестовой выборки.

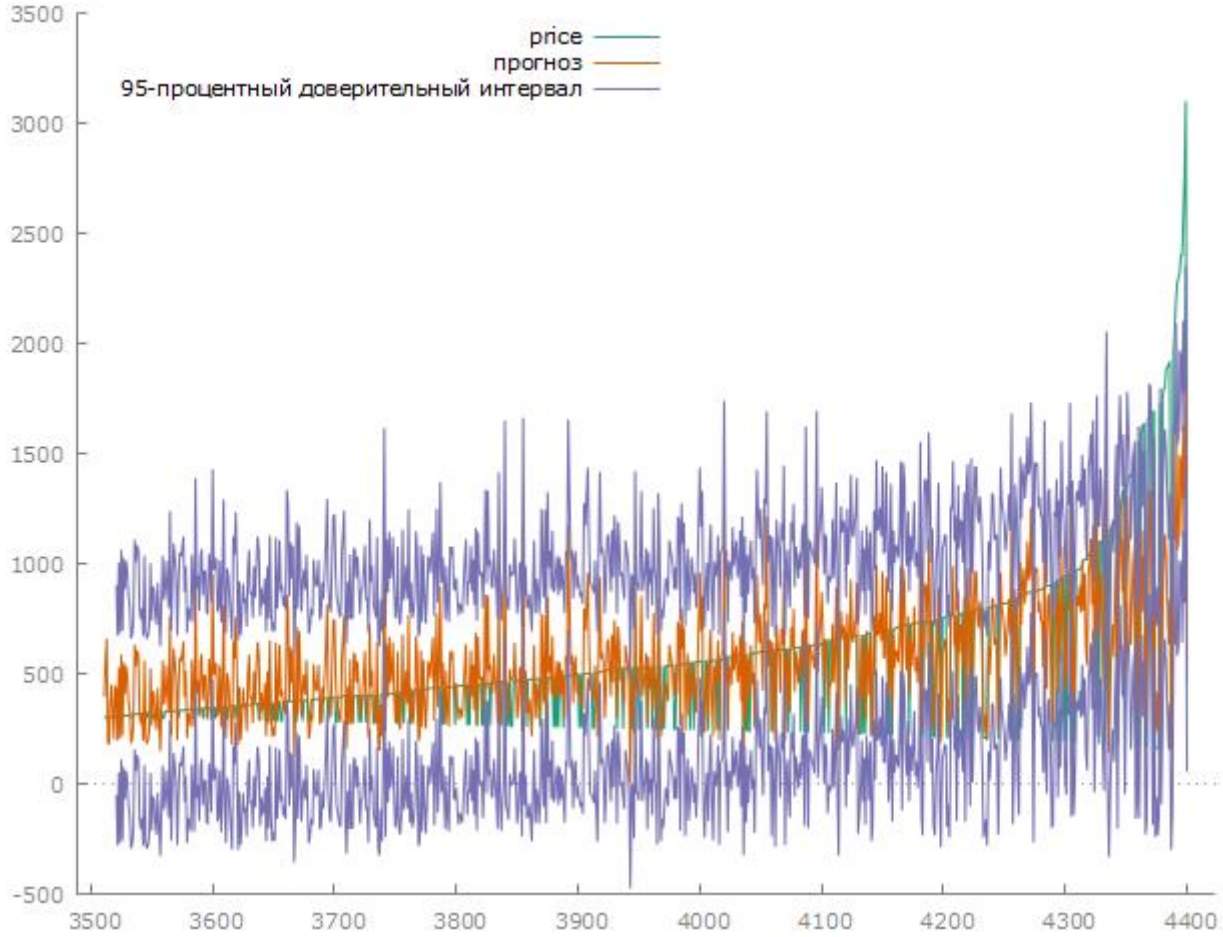

Рисунок 31. Прогнозы и истинные значения цены, 95-процентный доверительный интервал.

Было построено множество 95% доверительных интервалов для значений зависимой переменной и была посчитана среднеквадратическая погрешность, которая оказалась равной 252.14. Доля измерений, которые накрываются этими интервалами, была посчитана с помощью Python и равняется 94.43%.

4 Выводы и рекомендации

Теперь решим поставленную в 1.1 задачу. Определим *оптимальную стоимость загородного дома с заданными параметрами, а именно, многоэтажного дома без подвала возрастом не более 30 лет с жилой площадью не менее 2021 квадратных футов, площадью участка около 8500 квадратных футов, который не был на реконструкции.*

Согласно модифицированной модели 2 стоимость дома можно вычислять по следующей формуле:

$$price = 5005{,}76 + 0{,}295197 * sqft\_living - 0{,}00491793 * sqft\_lot\_ind * sqft\_lot - 0{,}00226988 * sqft\_lot - 29{,}3552 * floor - 0{,}00464971 * renovated * year\_built - 2{,}54046 * year\_built,$$

где $sqft\_lot\_ind = 1 * (sqft\_lot < 20000) + 0 * (sqft\_lot \geq 20000)$.

Так как в ходе анализа гипотез и модели выяснилось, что чем позже построен дом, тем он дешевле, то в слагаемых $-0{,}00464971 * renovated * year\_built$ и $-2{,}54046 * year\_built$ подставляем значение года постройки равное 2021, чтобы минимизировать итоговую цену.

Подставив все необходимые значения в формулу, получим значение цены дома с указанными параметрами:

$$price = 5005.76 + 0.295197 * 2021 - 0.00491793 * 8500 - 0.00226988 * 8500 - 29.3552 * 0 - 0.00464971 * 0 * 2021 - 2.54046 * 2021 = 406.987 \text{ тыс. долларов}.$$

Особенностью построенной модели можно считать то, что увеличение площади участка и увеличение года постройки (то есть уменьшение возраста) дома влияют на цену в сторону ее уменьшения, хотя изначально были выдвинуты предположения, утверждающие обратное. Если предварительный анализ данных не подтвердил лишь гипотезу 1, то модифицированные модели не подтвердили ни гипотезу 1, ни гипотезу 2.

Стоит также отметить, как потребитель может использовать полученные результаты для решения своих задач. Для того чтобы определить примерную стоимость своего дома на рынке, продавцу необходимо подставить значения параметров его дома в модель и тем самым определить приемлемую цену.