ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет экономических наук

Андреев Иван Васильевич
Андрюхин Борис Дмитриевич
Гусева Людмила
ЭАД

Проект по эконометрике

Москва 2024.

I. Введение.

1. Данные.

Данные о рынке недвижимости Сан-Паулу были импортированы с Kaggle. Они представляют собой парсинг с популярного в Бразилии ресурса (открывать с VPN) с объявлениями о продаже недвижимости. В датасете было около 25000 наблюдений, после очистки от выбросов осталось 15500.

В качестве целевой переменной была выбрана стоимость апартаментов. Независимыми признаками выступают Area (площадь), Adress (адрес), Bedrooms (количество спален), Bathrooms (количество ванных) и Parking_Spaces (количество парковочных мест). Также мы имеем дамми-переменную below_price, означающую нахождение под нижней границей средней стоимости. Однако она нам не подходит, так как напрямую зависит от цены апартаментов, позже заменим ее на другую. Автор парсера закодировал адреса координатами, что пригодится нам в будущем: в дальнейшем используем координаты для составления нового категориального признака, убрав из модели адрес апартаментов.

В процессе работы над проектом были приобщены дополнительные данные: координаты медицинских учреждений в Сан-Паулу, взятые с официального сайта префектуры Сан-Паулу (открывать с VPN), и координаты станций метро, взятые с Kaggle.

2. Цель исследования:

Цель эконометрического исследования состоит в изучении механизмов ценообразования на рынке недвижимости Сан-Паулу и выделении ключевых

факторов, влияющие на стоимость объектов недвижимости в разных районах города.

3. Задачи:

- Идентификация ключевых факторов ценообразования:

Систематизировать и проанализировать факторы, влияющие на стоимость недвижимости в Сан-Паулу, разделив их на группы (внешние и внутренние характеристики).

- Количественный анализ влияния факторов:

Применение статистических методов (регрессионный анализ, корреляционный анализ) для оценки и интерпретации количественного влияния выявленных факторов на ценообразование недвижимости. Это потребует сбора и обработки данных о ценах, характеристиках объектов.

- Сравнительный анализ цен в разных районах:

Сравнение цен на аналогичные объекты недвижимости в разных районах Сан-Паулу с учетом выявленных факторов, определение различий в ценообразовании и их причин. Картографический анализ может быть полезным инструментом для визуализации этих различий.

- Разработка модели ценообразования:

На основе полученных результатов разработка модели – множественной линейной регрессии, позволяющей прогнозировать стоимость недвижимости в Сан-Пауло с учетом ключевых факторов (в эконометрическом исследовании прогнозирование является сопутствующей задачей).

4. Актуальность исследования:

Рынок недвижимости Сан-Паулу является одним из крупнейших и наиболее динамичных в Латинской Америке. Его изучение крайне актуально в связи с:

• Высоким инвестиционным потенциалом: Сан-Паулу привлекает значительные объемы иностранных и внутренних инвестиций в недвижимость[1].

• Сложностью и многообразием рынка: Рынок недвижимости Сан-Паулу характеризуется высокой степенью сложности[2]

II. Экономическая модель.

1. Объясняющие переменные.

- Категориальные.

- Расположенность в центре (*дамми-переменная*). Мы выделили по координатам на карте примерный центр города, а затем сопоставили квартиры, которые находятся в нем со значением 1 и 0 в противном случае. Получили бинарный признак.

- Областные квадраты. Город мы разбили на сетку, затем, используя координаты объектов недвижимости, были сопоставлены квадраты разбиения и объекты, которые в них находятся. Затем с помощью one-hot-encoding кодирования получили n-ное количество дамми-переменных, каждая из которых отвечает за расположение апартаментов в данном квадрате.

- Числовые.
  - Спальни – количество спальных комнат в штуках.
  - Ванные – количество ванных комнат в штуках.
  - Парковочные места – количество парковочных мест в штуках.
  - Площадь – площадь апартаментов в квадратных метрах.

---

[1] POWER OF ATTRACTIVENESS OF BRAZILIAN CITIES FOR INTERNATIONAL REAL ESTATE INVESTMENTS: THE CASE OF CURITIBA Marzia Morena, Tommaso Truppi, Caio Smolarek Dias

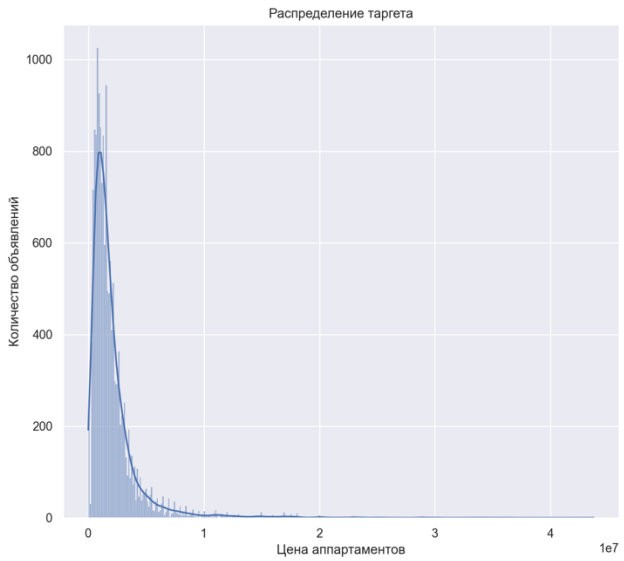
[2] DETERMINANTES DOS PREÇOS DE IMÓVEIS RESIDENCIAIS VERTICAIS NO MUNICÍPIO DE SÃO PAULO *Alexandre Esberard Gomes Vladimir Fernandes Maciel Mónica Yukie Kuwahara*

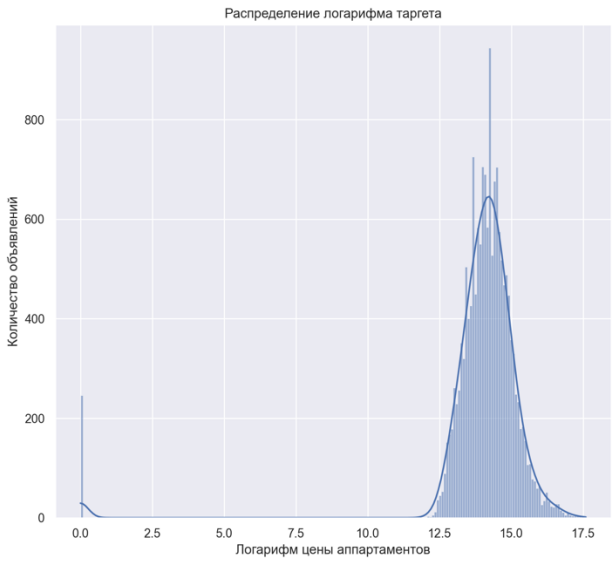
- Минимальное расстояние до метро – расстояние в километрах до ближайшей станции метро.
- Количество медицинских учреждений в квадрате – в штуках количество таких учреждений в определенных ранее областях.

Были выбраны именно эти объясняющие переменные из всего датасета, поскольку id – не имеет смысла для модели, а признак выше или ниже рыночной цены в датасете (below_price) может привести к переобучению за счет таргет ликеджа, но и он также не имеет смысла для интерпретации.

2. Целевая переменная.

Целевой переменной нашего исследования является цена апартаментов в Сан-Паулу и его ближайшем округе в бразильских реалах. Распределение объясняемой переменной не симметричное и имеет тяжелые хвосты справа. Логарифм цены апартаментов распределен более "нормально", поэтому обозначим его новым таргетом. При нормальности целевой переменной и остатков линейная регрессия дает несмещенные, эффективные и состоятельные оценки параметров (метод наименьших квадратов становится оптимальным). Это связано с тем, что метод МНК лучше работает в условиях нормального распределения.





3. Предположительное влияние переменных на таргет.

Расположенность в центре, количество спален, количество ванных комнат, площадь, количество парковочных мест – все числовые признаки будут иметь положительную корреляцию с целевой переменной. А расположенность в центре будет также будет увеличивать логарифм стоимости.

Количество медицинских учреждений в областном квадрате предположительно имеет положительную корреляцию с целевой переменной, социальные объекты увеличивают привлекательность района.

Расстояние от апартаментов до метро предположительно имеет отрицательную корреляцию с целевой переменной. Инфраструктура района оказывает большое влияние на его привлекательность.

Гипотезы:

- Влияние площади на цену может быть нелинейным. Рост цены на единицу площади может замедляться по мере увеличения размера объекта, отрицательная отдача от масштаба.
- Близость к метро будет отрицательно коррелировать с ценой недвижимости (расстояние меньше – цена больше), так как эти факторы повышают привлекательность местоположения за счет удобства транспортной системы.
- Местоположение играет роль. Нахождение в разных областных квадратах значимо.

4. *Дополнительные пояснения по работе, основанные на источниках.*

- Описание характеристик объекта недвижимости.

В качестве внутренних характеристик используются площадь, количество комнат, количество парковочных мест, количество ванных комнат, квадрат в котором находится объект.

В качестве внешних характеристик[3] используется расстояние до ближайшей станции метро и количество медицинских учреждений в квадрате (город был на них разбит предварительно).

Важно учитывать внешние факторы поскольку, согласно исследованию Luiz Paulo Lopes Fávero, они значимы, и их недооценка приведет к искажению в модели.

5. Характеристика районов

Как уже было сказано в актуальности работы, рынок недвижимости Сан-Пауло очень неоднородный, поскольку этот город является финансовым центром Бразилии, в нем немало престижных районов, однако бедные и преступные тоже присутствуют.

- Престижные районы:
  - Jardim Paulista.
  - Itaim Bibi.
  - Morumbi.
  - Pinheiros.
  - Higienópolis.
- Бедные районы:
  - Paraisópolis.
  - Heliópolis.
  - Guaianases

[3] MODELOS DE PREÇOS HEDÔNICOS APLICADOS A IMÓVEIS RESIDENCIAIS EM LANÇAMENTO NO MUNICÍPIO DE SÃO PAULO Luiz Paulo Lopes Fávero

- Perus.

В разных районах может быть своя структура принятия решений относительно покупки недвижимости, например, в некоторых районах спрос на новое жилье намного выше, чем предложение, в то время как в других районах наблюдается обратная ситуация, при этом в Сан-Паулу предложение новых жилых объектов в городе недостаточно для удовлетворения текущего и будущего спроса[4], что может завышать общий уровень цен.

В работе Alexandre Esberard Gomes, Vladimir Fernandes Maciel, Mônica Yukie Kuwahara[5] много полезных графиков и диаграмм, которые помогают лучше понять структуру Сан-Паулу: где находятся деловые и промышленные центры, как работает транспортная система города, распределение рабочих мест и так далее. Благодаря статье мы и решили включить станции метро и медицинские учреждения в наше исследование.

III. Предварительный анализ данных.

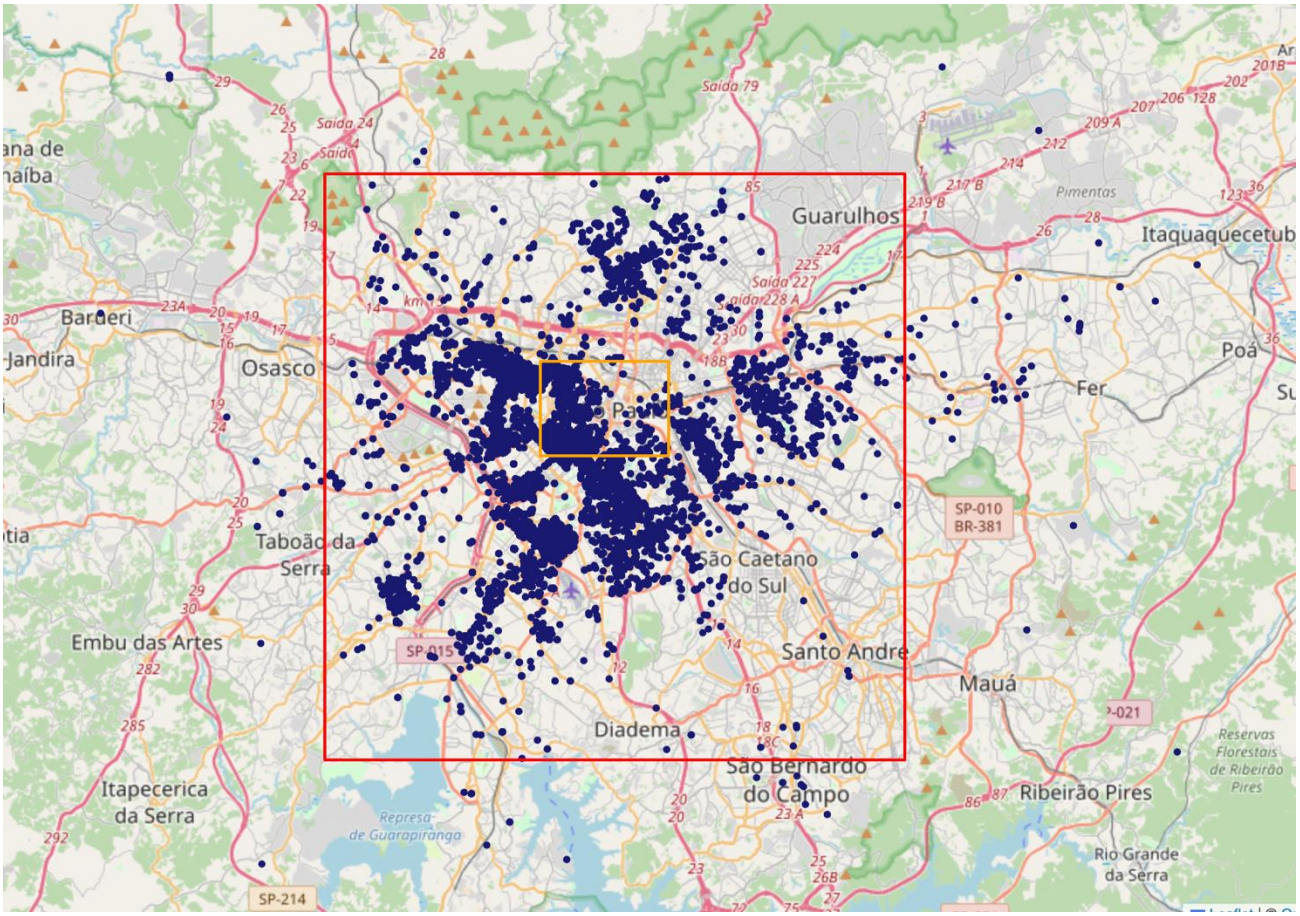
1. Обработка данных.

Важнейшим этапом в анализе данных является обработка этих самых данных. Сначала посмотрим пропуски в датасете, чтобы получить максимально возможную точность модели. Автор парсера попытался закодировать адреса

[4] Tendências Imobiliárias: análise da demanda e da oferta por imóveis residenciais em São Paulo utilizando survey e dados secundários.
Real Estate Trends: analysis of demand and supply for housing in Sao Paulo using survey and secondary data.
João Francisco Resende, Rosi Rosendo

[5] DETERMINANTES DOS PREÇOS DE IMÓVEIS RESIDENCIAIS VERTICAIS NO MUNICÍPIO DE SÃO PAULO *Alexandre Esberard Gomes Vladimir Fernandes Maciel Mônica Yukie Kuwahara*

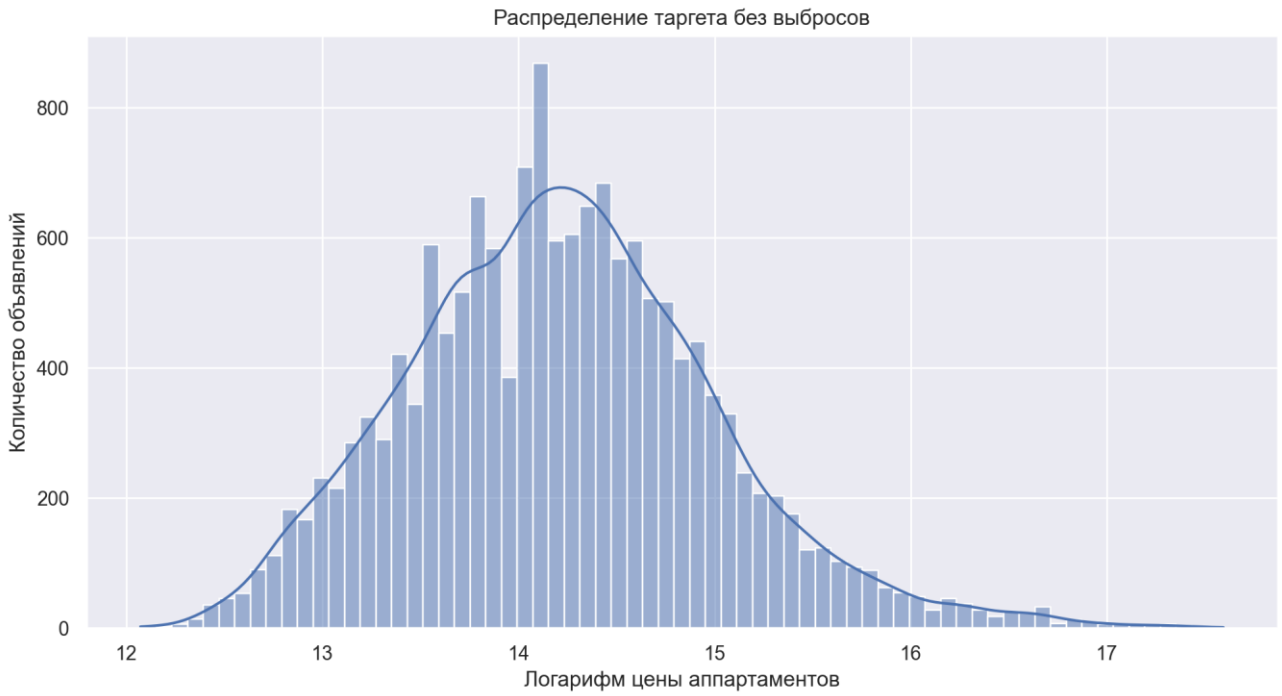
координатами, но, к сожалению, почти треть из них отсутствует. Посмотрим на распределение координат без пропусков.



Заметим, что город плотно покрыт точками. Можем удалить из данных наблюдения с пропусками. Также можно заметить выбросы: в датасете находятся апартаменты, находящиеся далеко за границами Сан-Паулу. Оставим в датасете строки с координатами, входящими в красный квадрат: Сан-Паулу и ближайший пригород.

Определим дамми-переменную расположенность в центре Сан-Паулу. В нее войдут точки, находящиеся в оранжевом квадрате. Позже мы увидим, как это влияет на цену апартаментов.

Пропусков в данных больше нет. Посмотрим на распределение признаков. Вспомните, на графике распределения таргета было видно большое количество выбросов. Исправим это.



С целевой переменной разобрались. Посмотрим на распределение категориальных признаков. Но для начала определим количество уникальных значений для каждого из них.

unique Center - 2

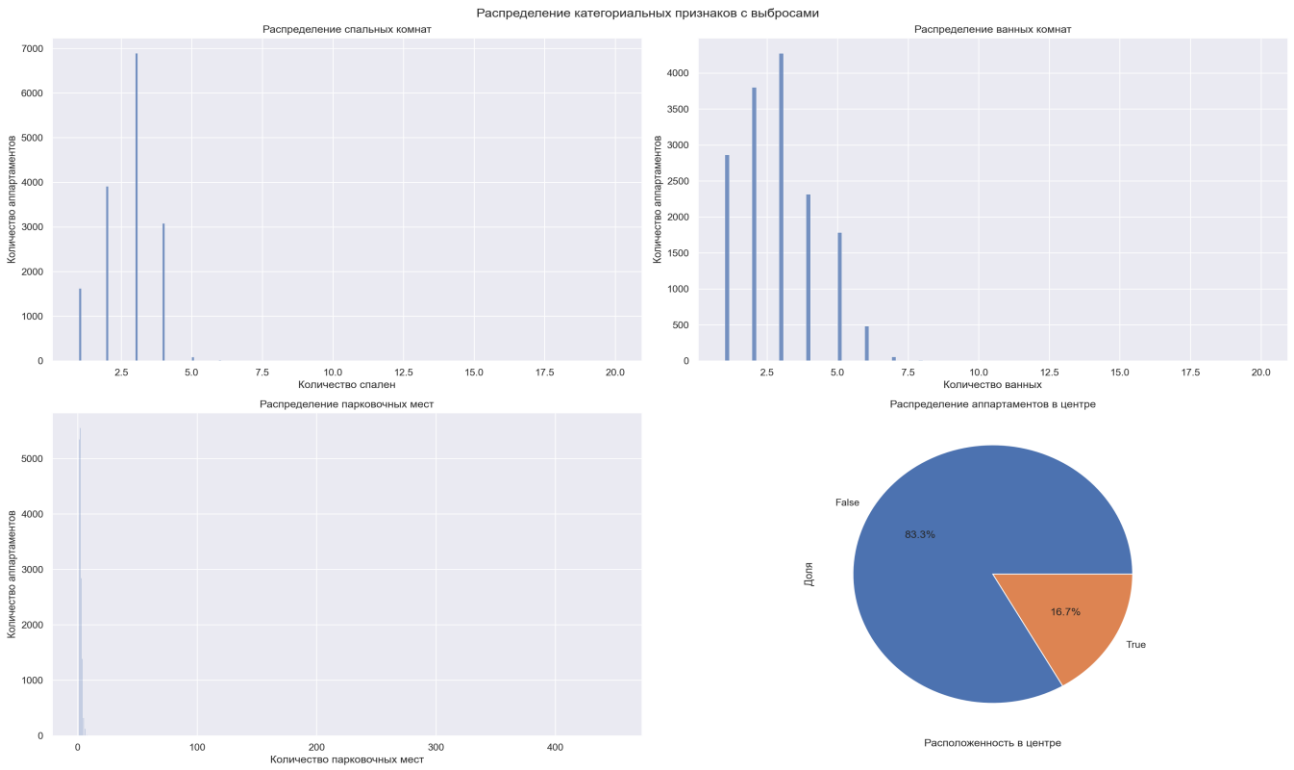
unique Adress - 8386

unique Bedrooms - 10

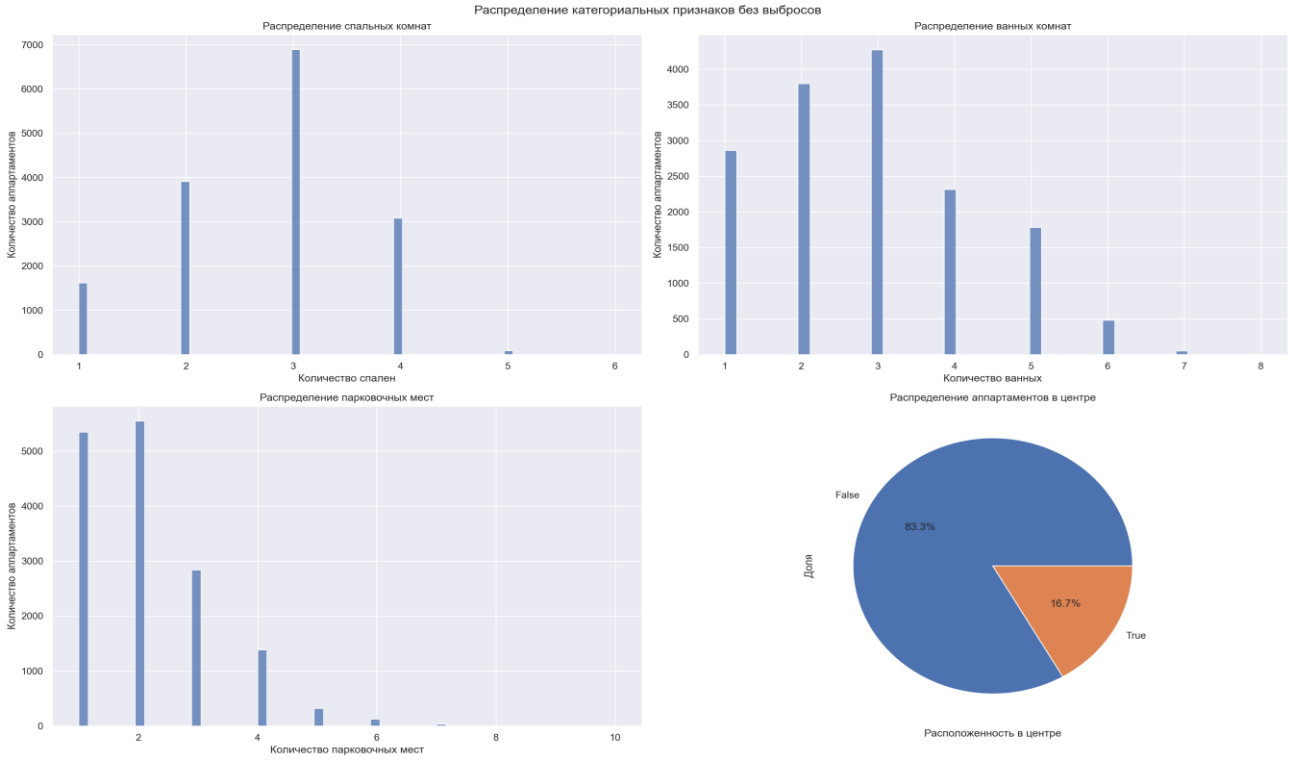
unique Bathrooms - 12

unique Parking_Spaces - 20

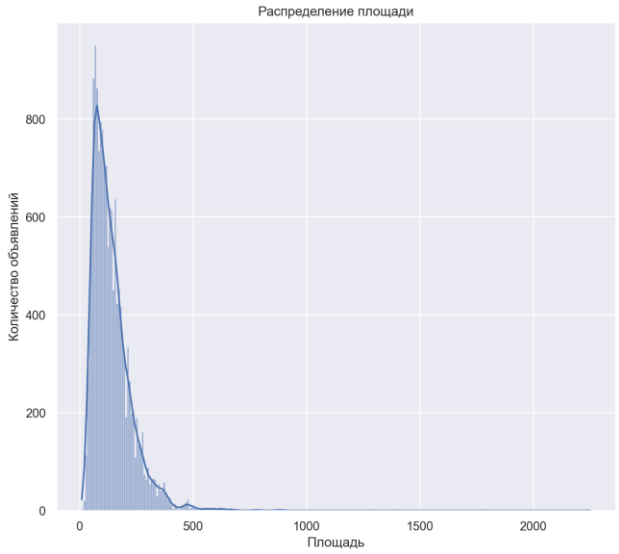
Построим графики распределений для категориальных признаков с адекватным количеством уникальных значений и обработаем выбросы.

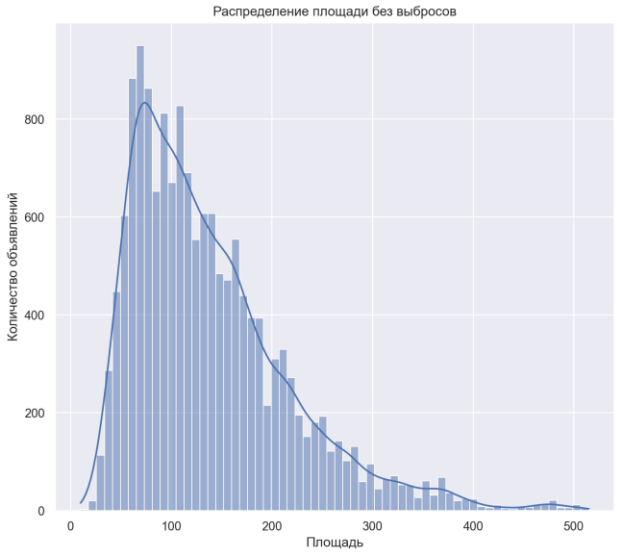


Выбросов много. Разберемся с ними и снова построим графики.



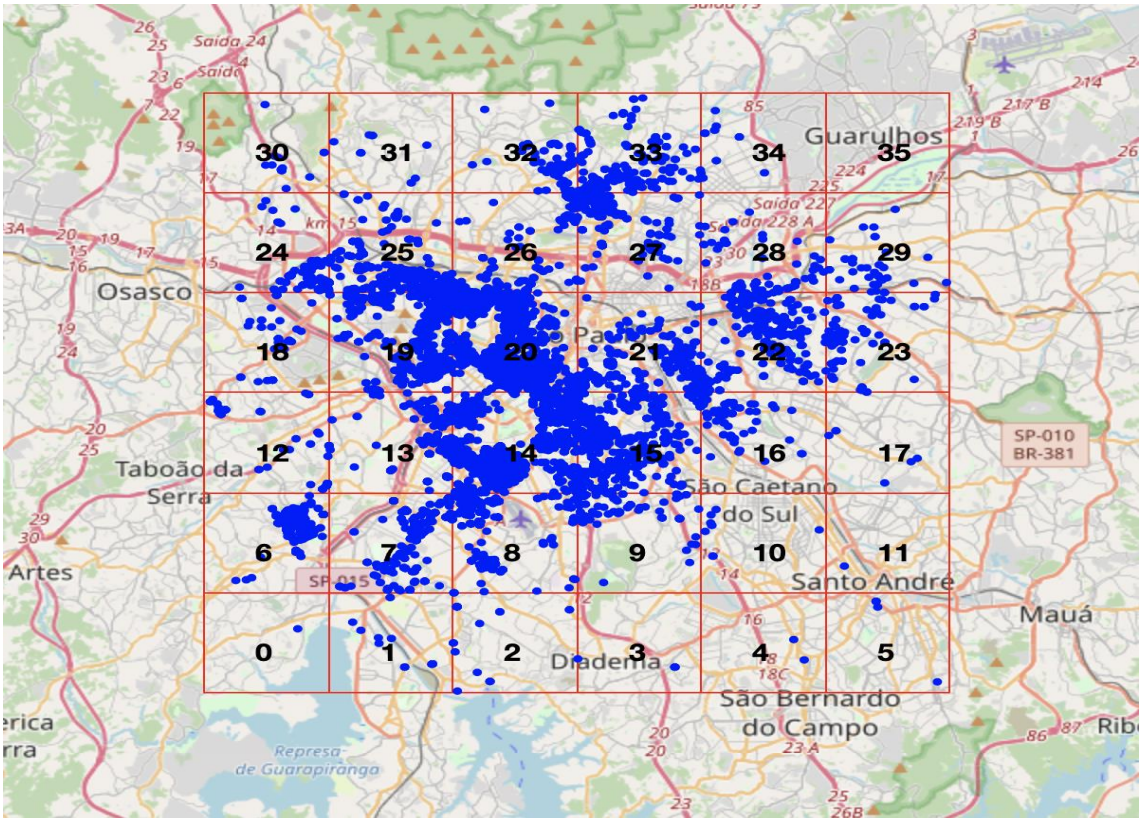
Так мы избавились от выбросов в категориальных признаках. Посмотрим на распределение единственного числового признака - площадь.



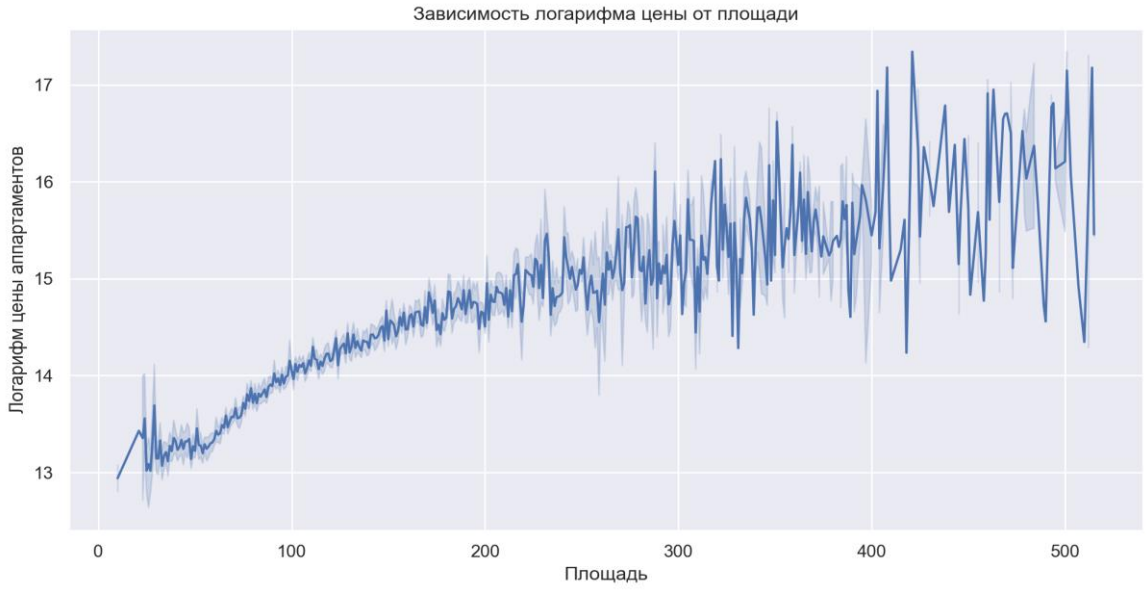


Опять видим тяжелые хвосты справа. Отрежем с помощью квантили и избавимся от выбросов.

Но что делать с адресами? Слишком большое количество уникальных значений, хотя признак действительно кажется важным. Для этого нам и нужны были координаты. Разобьем красный квадрат на сетку и каждому наблюдению присвоим номер сетки, куда он попал.

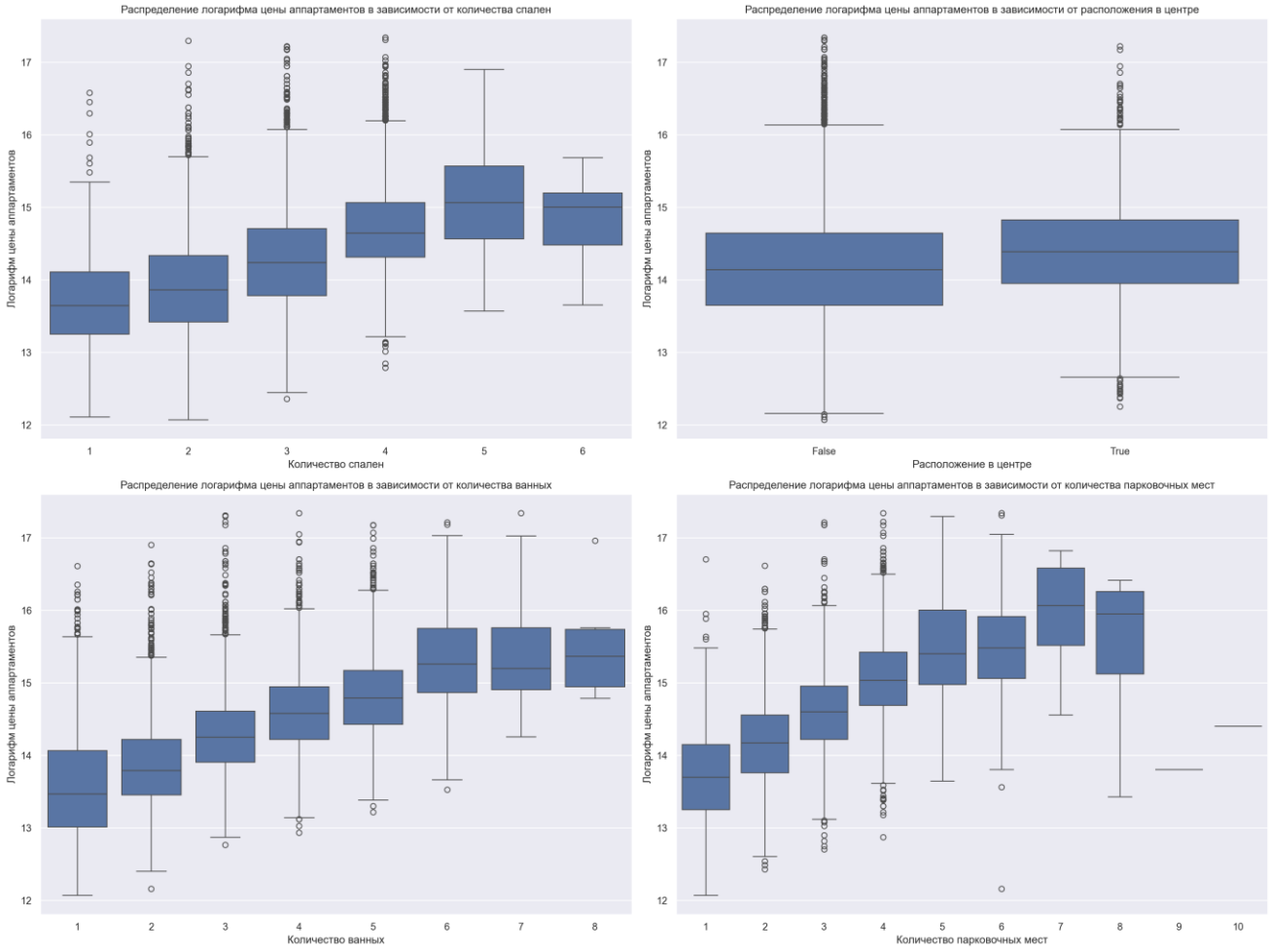


Давайте наконец построим графики зависимости логарифма цены апартаментов от независимых признаков. Начнем с числового признака - площади.



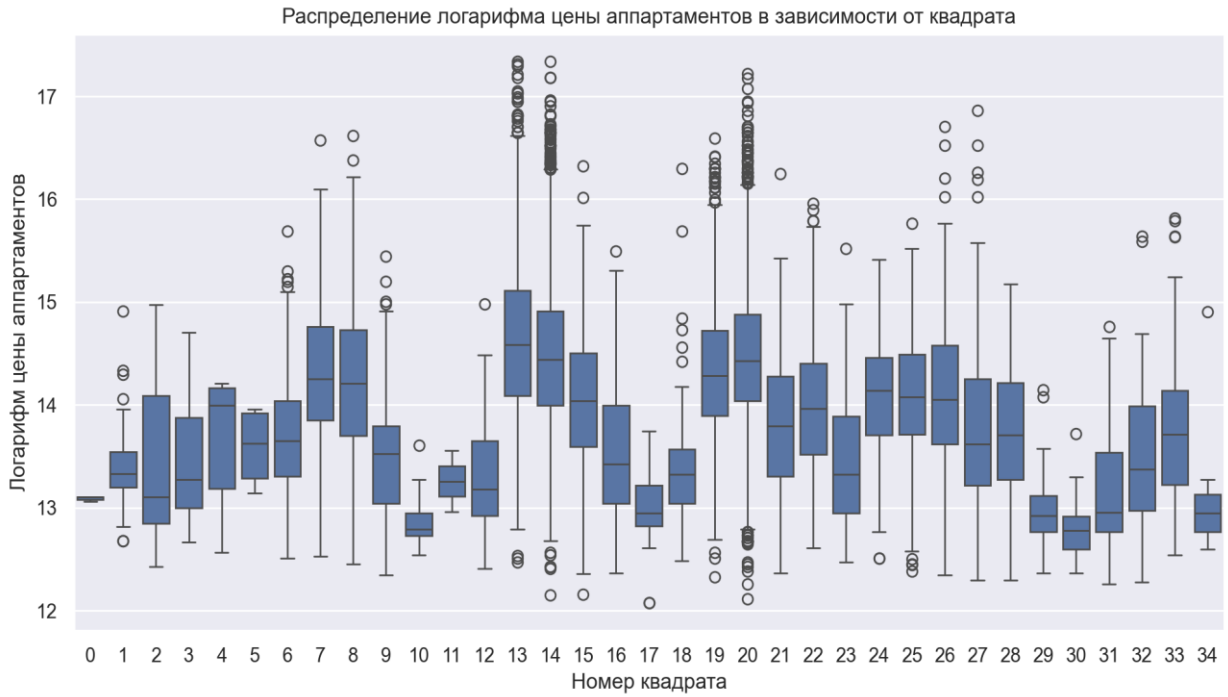
Наблюдается сильная положительная линейная связь, о чем нам также говорит корреляция площади и цены на апартаменты. Однако по мере увеличения площади рост цен замедляется и график становится шумным.

Перейдем к категориальным признакам. Для них изобразим ящики с усами.

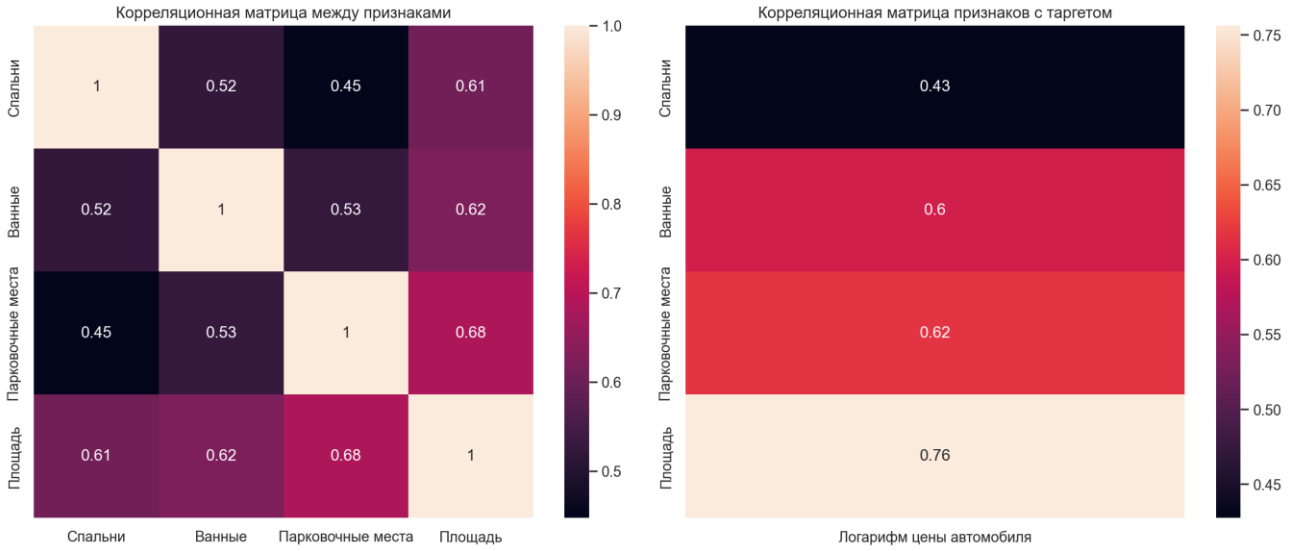


Кажется, ванные, спальни и парковочные места имеют что-то похожее на линейную зависимость. Может, сделаем эти признаки числовыми?

Посмотрим, что дало нам разделение города на квадраты. Данные дамми-переменные окажутся весьма полезными. На графике видна большая разрозненность.



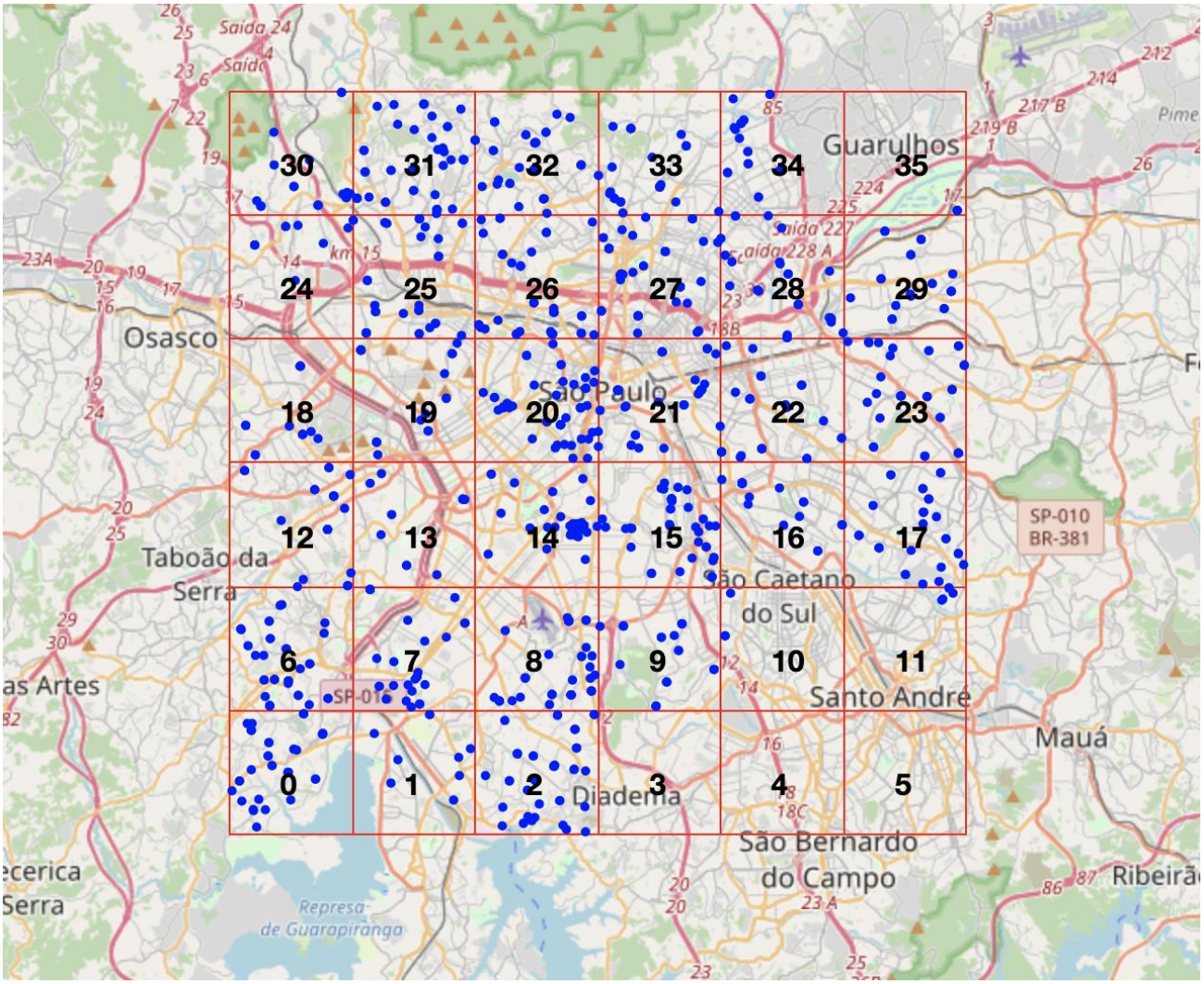
Давайте также построим корреляционную матрицу.



У признаков наблюдается средняя положительная корреляция между друг другом, что, в целом, логично: например, чем больше площадь, тем больше спален. Это плохо отразится на точности модели, но что поделать.

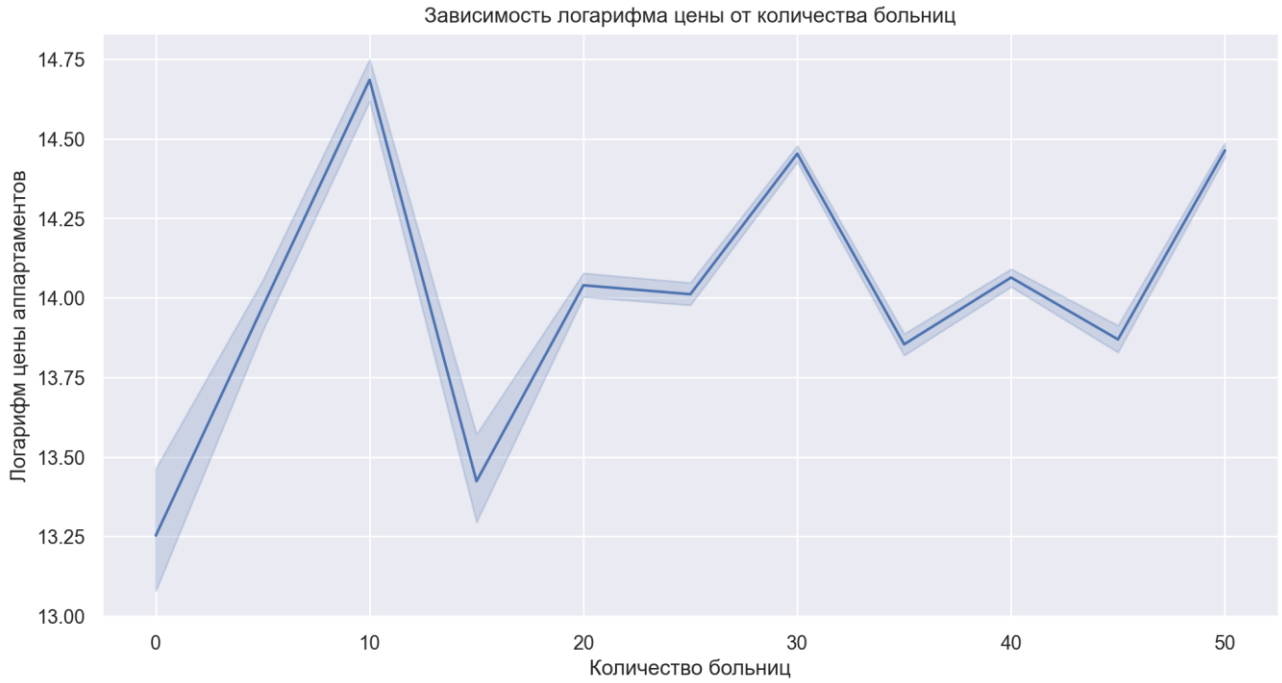
Между признаками и таргетом так же наблюдается средняя и сильная положительная корреляция, что хорошо скажется на точности модели.

Осталось реализовать еще одну идею. На официальном сайте префектуры Сан-Паулу мы нашли датасет с медицинскими учреждениями и их координатами. Влияет ли близость апартаментов к больницам на их цену? Посмотрим на график распределения учреждений по квадратам.



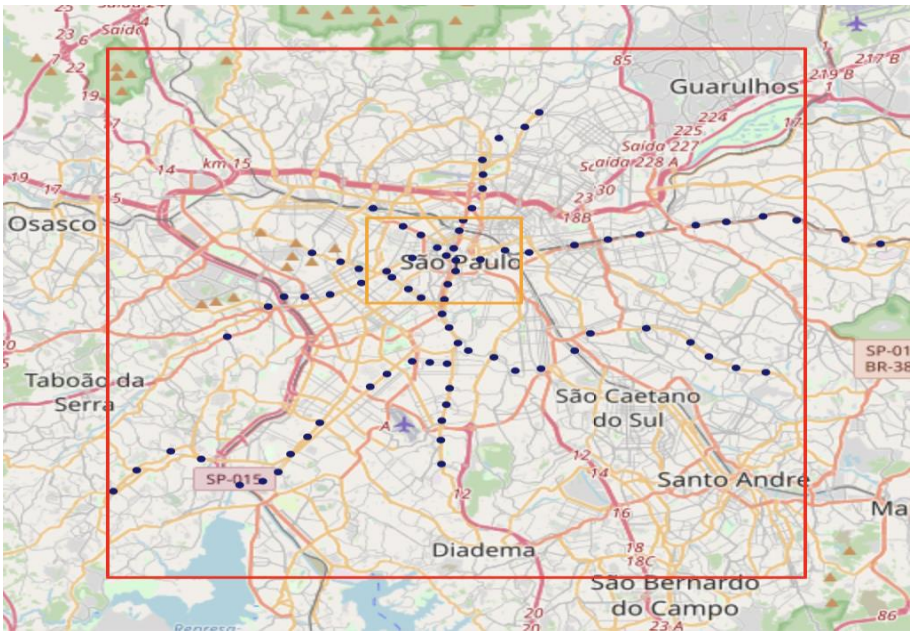
Посмотрим на зависимость таргета от нового признака.

Корреляция количества больниц в районе и цены на аппартаменты: 0.05092009031365304



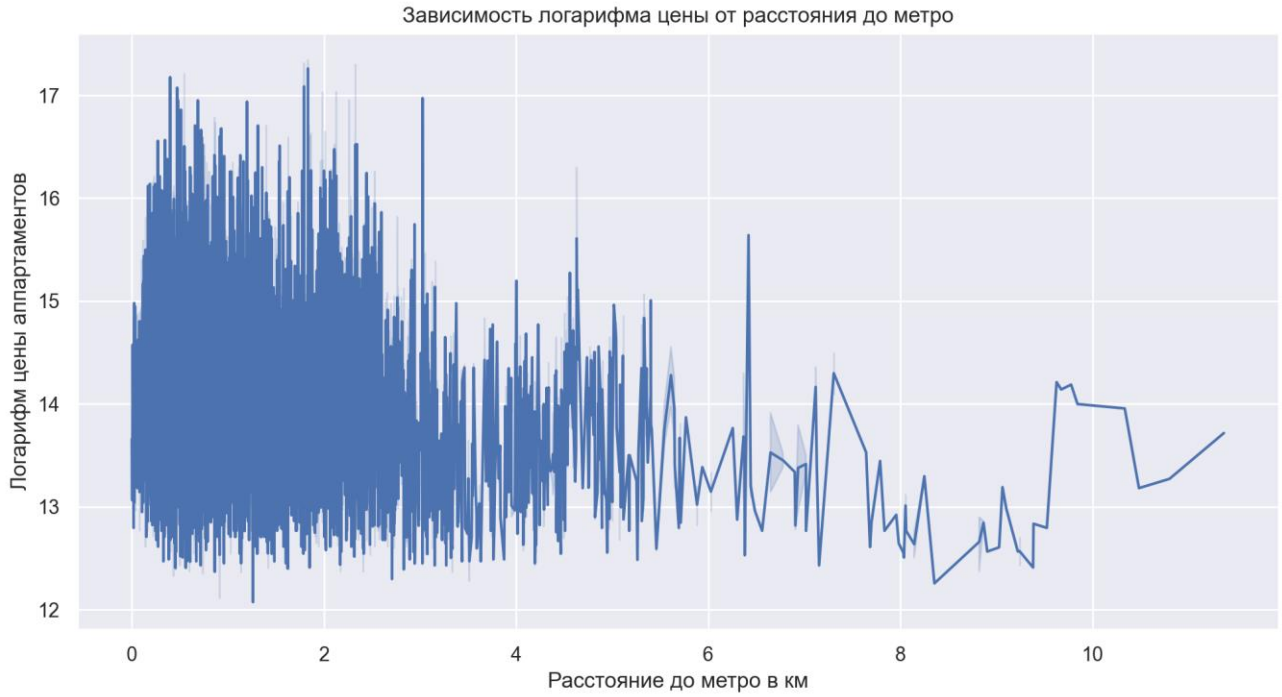
Конечно, корреляция вышла не такая большая и график не слишком линейным. Однако переменная, может, действительно оказаться значимой.

Рассмотрим еще один новый признак - расстояние до метро.



Посчитаем расстояние в километрах от апартаментов до каждой станции метро и найдем минимальное расстояние. Посмотрим на корреляцию нового признака и таргета и построим график зависимости.

Корреляция расстояния до метро и цены на апартаменты: -0.16940099950040577



Видим слабую отрицательную корреляцию расстояния до метро и логарифма цены апартаментов и очень шумный график. Посмотрим, что нам даст обучение.

IV. Оценка модели.

Наконец, мы можем перейти к обучению. Но для начала данные следует закодировать. Для категориальных признаков (квадраты координат) воспользуемся OneHotEncoder-ом - создадим еще n-ное количество дамми-переменных, каждая из которых отвечает за расположение аппартаментов в данном квадрате. Для числовых переменных (спальни, ванные, парковочные

места, площадь, количество больниц в квадрате, расстояние до метро) воспользуемся StandardScaler-ом - отнормируем переменные. Обучим множественную линейную регрессию и посмотрим на результаты. Результаты выгрузили в красивую таблицу с использованием библиотек pandas и docx.

Результаты модели множественной линейной регрессии

Parameter	Estimate	Std Error	t-Statistic	P-Value
const	13.4067	0.0130	1028.1351	0.0000
ohe__Coord_cell_0	0.3717	0.2262	1.6432	0.1004
ohe__Coord_cell_1	0.0786	0.0731	1.0761	0.2819
ohe__Coord_cell_2	0.6164	0.0664	9.2813	0.0000
ohe__Coord_cell_3	0.0784	0.1353	0.5791	0.5625
ohe__Coord_cell_4	0.3223	0.1457	2.2123	0.0270
ohe__Coord_cell_5	-0.1325	0.1867	-0.7094	0.4781
ohe__Coord_cell_6	0.3358	0.0213	15.7789	0.0000
ohe__Coord_cell_7	0.6432	0.0250	25.7508	0.0000
ohe__Coord_cell_8	0.8050	0.0215	37.4525	0.0000
ohe__Coord_cell_9	0.1166	0.0372	3.1355	0.0017
ohe__Coord_cell_10	-0.4241	0.1008	-4.2065	0.0000
ohe__Coord_cell_11	-0.1244	0.2678	-0.4643	0.6424
ohe__Coord_cell_12	0.0338	0.0455	0.7444	0.4566
ohe__Coord_cell_13	0.8432	0.0362	23.2745	0.0000
ohe__Coord_cell_14	0.9700	0.0165	58.7519	0.0000
ohe__Coord_cell_15	0.7593	0.0195	38.8676	0.0000
ohe__Coord_cell_16	0.2191	0.0412	5.3202	0.0000
ohe__Coord_cell_17	0.0014	0.0920	0.0154	0.9877
ohe__Coord_cell_18	0.1826	0.0511	3.5738	0.0004
ohe__Coord_cell_19	0.7117	0.0259	27.5180	0.0000
ohe__Coord_cell_20	1.2159	0.0285	42.7106	0.0000
ohe__Coord_cell_21	0.7120	0.0263	27.1163	0.0000
ohe__Coord_cell_22	0.3953	0.0237	16.7019	0.0000
ohe__Coord_cell_23	0.4447	0.0339	13.1120	0.0000
ohe__Coord_cell_24	0.4098	0.0466	8.8037	0.0000
ohe__Coord_cell_25	0.8302	0.0202	41.0655	0.0000
ohe__Coord_cell_26	0.8881	0.0287	30.9939	0.0000
ohe__Coord_cell_27	0.5610	0.0270	20.7763	0.0000
ohe__Coord_cell_28	0.4720	0.0323	14.6069	0.0000
ohe__Coord_cell_29	-0.1088	0.0499	-2.1799	0.0293

ohe__Coord_cell_30	0.0962	0.0914	1.0518	0.2929
ohe__Coord_cell_31	0.6297	0.0753	8.3584	0.0000
ohe__Coord_cell_32	0.2571	0.0347	7.4201	0.0000
ohe__Coord_cell_33	0.1898	0.0277	6.8589	0.0000
ohe__Coord_cell_34	0.0055	0.0849	0.0645	0.9486
scaling__Bedrooms	-0.0221	0.0044	-5.0620	0.0000
scaling__Bathrooms	0.1356	0.0044	30.6514	0.0000
scaling__Parking_Spaces	0.1745	0.0049	35.5328	0.0000
scaling__Area	0.3431	0.0059	58.3755	0.0000
scaling__meds_count_group	-0.1998	0.0167	-11.9470	0.0000
scaling__metro_distance	-0.0375	0.0057	-6.6327	0.0000
already_ohe__Center	-0.0363	0.0137	-2.6444	0.0082

R-squared: 0.7187
Adjusted R-squared: 0.7180
F-statistic: 987.5404
Prob (F-statistic): 0.0000e+00
Log-Likelihood: -8086.7730

Интерпретируем результаты.

Общая характеристика модели:

- Коэффициент детерминации (R-squared): 0.719 — модель объясняет 71,9% вариации зависимой переменной.
- Скорректированный R-squared (Adj. R-squared): 0.718 — скорректированная версия R^2, учитывающая количество предикторов.
- F-статистика: 987.5 с вероятностью (Prob (F-statistic)) 0.00 — модель статистически значима.

Оценка коэффициентов

Каждый коэффициент (coef) показывает изменение логарифма цены при изменении соответствующего предиктора на одну единицу, при условии фиксированных остальных переменных. Рассмотрим основные группы предикторов:

1. Постоянная:

Const = 13.4067

Постоянная указывает базовый уровень логарифма цены, когда все предикторы равны нулю.

2. Двоичные переменные (One-Hot Encoding) для координатных ячеек (ohe__Coord_cell_):

Эти переменные представляют категориальные данные, связанные с расположением объекта в координатной сетке.

Значимые переменные: Большинство из них статистически значимы ($p < 0.05$), что указывает на существенное влияние конкретных координатных ячеек на цену.

Примеры:

ohe__Coord_cell_10 (клетка номер 10) имеет отрицательный коэффициент (-0.4241, $p < 0.001$), указывая на снижение цены для этой ячейки, а эта ячейка соответствует бедному району – Helióрolis.

А в клетках номер 13, 14, 19, 20, в которых находятся вышеописанные престижные районы, наоборот имеют положительные коэффициенты (0.8432, 0.9700, 0.7117, 1.2159, соответственно, при $p < 0.001$)

3. Отмасштабированные числовые переменные (scaling):

scaling__Bedrooms = -0.0221 ($p < 0.001$)

Следовательно, при увеличении на одну масштабированную единицу количества спален - уменьшается логарифм цены на 0.0221. То есть увеличение числа спален не всегда приводит к росту цены, возможно, из-за других факторов, влияющих на стоимость.

scaling__Bathrooms = 0.1356 ($p < 0.001$)

Увеличение числа ванных комнат на одну единицу увеличивает логарифм цены на 0.1356. Это логично, так как наличие большего числа ванных комнат повышает комфортабельность жилья.

scaling__Parking_Spaces = 0.1745 ($p < 0.001$)

Увеличение количества парковочных мест положительно влияет на цену, увеличивая её логарифм на 0.1745, что логично особенно для стран, в которых личные автомобили – важный атрибут повседневной жизни, а Бразилия является таковой.

scaling__Area = 0.3431 ($p < 0.001$)

Увеличение площади на одну масштабированную единицу увеличивает логарифм цены на 0.3431. Такое значительное влияние объясняется тем, что цена квадратного метра для разных типов объектов формируется на рынке, поэтому с увеличением площади, недвижимость должна дорожать.

scaling__meds_count_group = -0.1998 ($p < 0.001$)

Эта переменная связана с количеством медицинских учреждений в областном квадрате. Отрицательный коэффициент указывает на снижение цены при увеличении этого показателя, однако корреляция с логарифмом цены была положительная, хоть и слабая. Требуются дополнительные исследования этого показателя.

scaling__metro_distance = -0.0375 ($p < 0.001$)

Увеличение расстояния до метро связано с уменьшением цены, что логично, поскольку близость к метро обычно повышает стоимость жилья, так как инфраструктура района становится удобнее.

V. Выводы.

Первая гипотеза подтвердилась, глядя на график зависимости логарифма цены от площади.

Вторая гипотеза подтвердилась. Расстояние до ближайшей станции метро имеет отрицательную корреляцию с таргетом и отрицательный вес в модели, признак является значимым, исходя из описания модели.

Третья гипотеза подтвердилась. Действительно, большинство признаков оказались значимыми, а также особенно интересно, что в престижных и бедных районах коэффициенты сильно влияют на таргет.

VI. Источники.

1. Gomes A. E., Maciel V. F., Kuwahara M. Y. Determinantes dos preços de imóveis residenciais verticais no município de São Paulo //Anais do XL Encontro Nacional de Economia. – 2012. – С. 1-19.
2. Resende J. F., Rosendo R. Tendências Imobiliárias: análise da demanda e da oferta por imóveis residenciais em São Paulo utilizando survey e dados secundários. – Latin American Real Estate Society (LARES), 2009. – №. lares2009_157-305-1-rv.
3. Fávero L. P. L. Modelos de preços hedônicos aplicados a imóveis residenciais em lançamento no município de São Paulo : дис. – Universidade de São Paulo, 2003.
4. SMOLAREK DIAS C. Power of attractiveness of brazilian cities for international real estate investments: the case of Curitiba. – 2009.