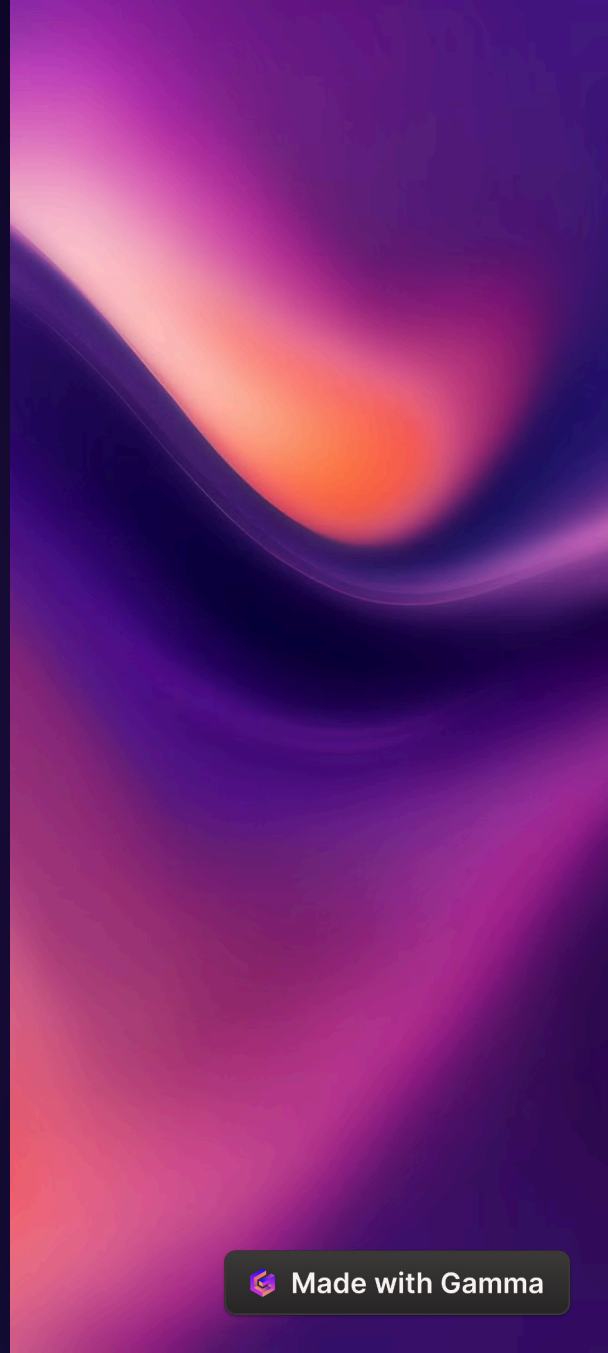
Исследование данных о сделках с недвижимостью

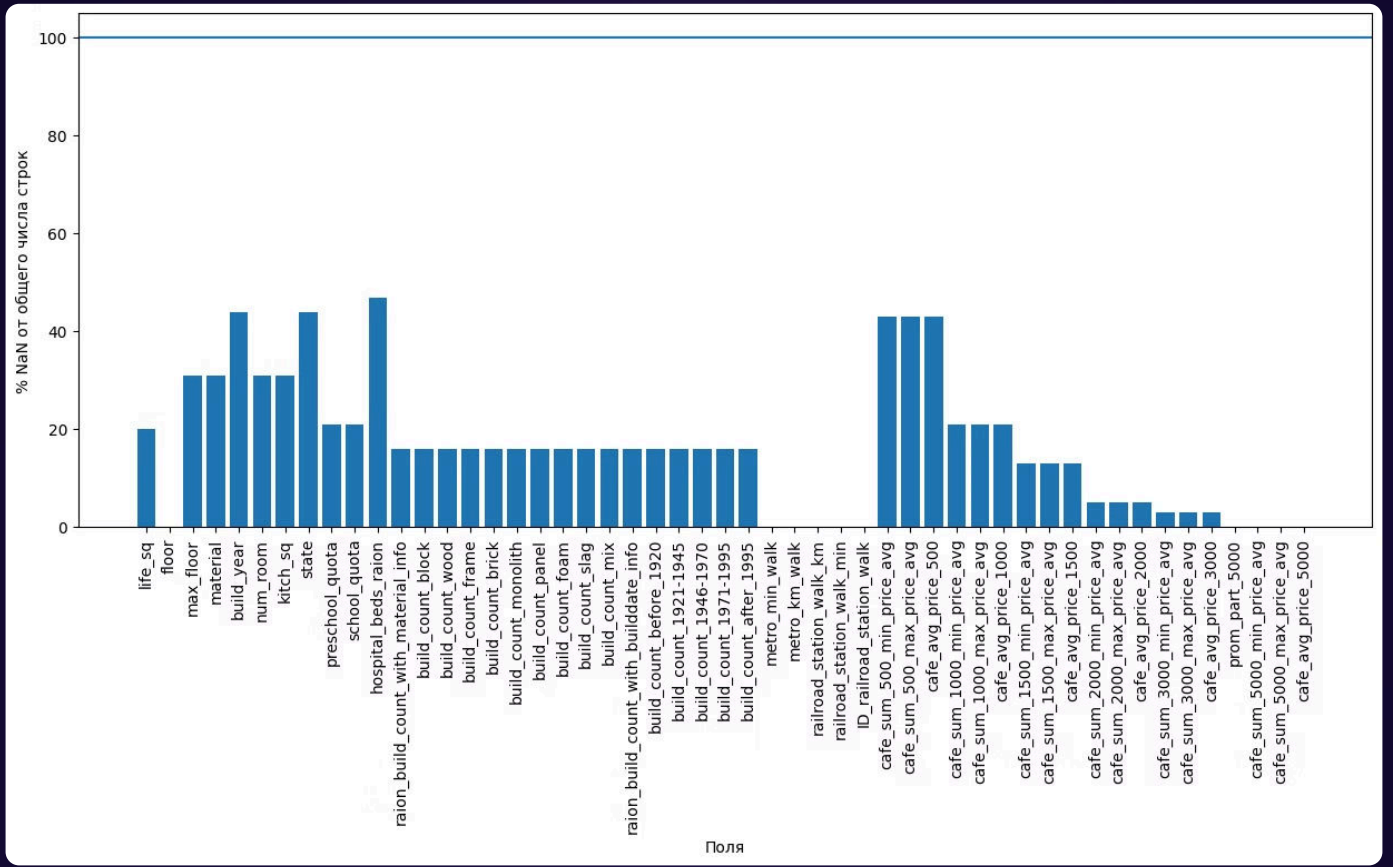
План

1. Обработка отсутствующих значений
2. Обработка лишних значений
3. Выявление аномалий
4. Сбалансированность данных
5. Базовый отбор признаков
6. Статистики



Обработка отсутствующих значений

Составим график, отображающий отношение отсутствующих и нормальных значений для каждого поля датасета:



Отсутствующие значения у полей *'cafe_sum_500_min_price_avg', 'cafe_sum_500_max_price_avg', 'cafe_avg_price_500', 'cafe_sum_1000_min_price_avg', 'cafe_sum_1000_max_price_avg', 'cafe_avg_price_1000', 'cafe_sum_1500_min_price_avg', 'cafe_sum_1500_max_price_avg', 'cafe_avg_price_1500', 'cafe_sum_2000_min_price_avg', 'cafe_sum_2000_max_price_avg', 'cafe_avg_price_2000', 'cafe_sum_3000_min_price_avg', 'cafe_sum_3000_max_price_avg', 'cafe_avg_price_3000'* можно спокойно заменить на средние по столбцу значения, так как, в целом, средний чек в кафе не будет сильно разнится в пределах 1 города.

'life_sq' (жилую площадь) можно считать, домножая общую площадь на коэффициент, равный среднему по всем данным отношению жилой к общей площади.

'max_floor', 'material', 'build_year', 'num_room' довольно важные параметры, при этом часто отсутствующие. Их довольно проблематично усреднять. Например, квартира, находящаяся на 3 этаже, может располагаться как в 5-и, так и в 25-и этажном здании. Следовательно, стоит провести дополнительный сбор данных. В крайнем случае, 'max_floor' можно приравнять к этажу квартиры, 'material' брать как моду, 'build_year' - среднее значение по остальным данным, 'num_room' - предсказывать по жилой площади.

'kitch_sq' (площадь кухни) также не будет критично различаться у разных объектов недвижимости, поэтому разумно заполнять пропуски средним значением.

'state' (состояние квартиры) возможно только усреднять, т.к. по другому такие данные не получить.

'preschool_quota' и 'school_quota' - кол-во мест в образовательных учреждениях примерно одинаково для районов одного города, допустимо брать средние значения по полю.

'hospital_beds_raion' (количество больничных коек для района) можно считать, воспользовавшись macro.csv: 'raion_popul' / **'hospital_beds_available_per_cap'** (соединяем данные по полю timestamp)

'raion_build_count_with_material_info', 'build_count_block', 'build_count_wood', 'build_count_frame', 'build_count_brick', 'build_count_monolith', 'build_count_panel', 'build_count_foam', 'build_count_slag', 'build_count_mix', 'raion_build_count_with_builddate_info', 'build_count_before_1920', 'build_count_1921-1945', 'build_count_1946-1970', 'build_count_1971-1995', 'build_count_after_1995' - данные, которые довольно трудно восстановить путём парсинга каких-либо ресурсов. По причине взаимосвязи друг с другом, их не следует каким либо образом усреднять. Лучшее решение - удаление строк с пропусками этих полей.

*под усреднением подразумевается взятие медианы, что поможет защитить от аномалий

Обработка лишних значений

Создадим корреляционную матрицу на основе обработанного от пропущенных значений датасета. Удалим одно из полей, чья корреляция с другим по модулю превышает 0.96(подобранное значение).

Удалённые столбцы:

'cafe_avg_price_5000', 'cafe_sum_500_max_price_avg', '0_6_female', '16_29_male', 'cafe_count_2000_price_2500', '0_13_female', 'cafe_count_1000_price_2500', 'cafe_count_1000', 'office_count_1000', 'cafe_count_2000_price_4000', 'cafe_count_5000_price_1500', 'big_church_count_3000', 'cafe_count_3000_price_high', 'public_transport_station_km', 'ekder_male', 'school_km', 'church_count_1000', 'cafe_count_500_price_500', 'cafe_count_2000_price_1000', 'cafe_count_1500_price_2500', '0_6_male', 'cafe_sum_5000_min_price_avg', 'big_church_count_1500', 'cafe_avg_price_500', 'cafe_count_3000_price_500', 'cafe_avg_price_1000', 'church_count_2000', 'raion_build_count_with_builddate_info', 'cafe_count_5000_price_500', 'cafe_sum_2000_min_price_avg', 'female_f', '7_14_female', 'office_count_3000', 'cafe_count_500', 'cafe_count_5000', 'sadovoe_km', 'office_count_2000', '0_17_female', 'cafe_avg_price_1500', 'cafe_count_1500_price_1000', 'cafe_sum_1500_max_price_avg', '16_29_all', 'cafe_count_2000', 'cafe_count_1500_price_1500', 'cafe_count_1000_price_4000', 'bulvar_ring_km', 'railroad_station_walk_min', 'cafe_count_3000', 'office_count_1500', '0_13_all', 'cafe_count_3000_price_1000', 'raion_popul', 'cafe_count_3000_price_4000', 'cafe_sum_3000_min_price_avg', 'leisure_count_5000', 'cafe_count_3000_price_2500', 'cafe_count_2000_na_price', 'cafe_count_1500_price_4000', 'metro_km_walk', 'full_all', 'cafe_count_2000_price_1500', 'cafe_count_1500', 'cafe_count_5000_price_high', 'children_school', 'cafe_count_5000_price_1000', 'cafe_count_1500_price_500', 'big_church_count_1000', 'cafe_count_2000_price_high', 'young_all', 'leisure_count_1500', 'church_count_5000', 'metro_km_avto', 'work_female', 'cafe_count_1500_na_price', 'cafe_count_1000_price_1000', 'cafe_count_3000_na_price', 'cafe_count_1000_na_price', 'leisure_count_2000', 'cafe_count_1000_price_1500', 'cafe_count_5000_price_4000', 'church_count_1500', 'cafe_count_1000_price_500', 'office_count_5000', 'office_sqm_5000', 'young_female', 'big_church_count_2000', 'church_count_3000', 'cafe_avg_price_2000', 'male_f', 'cafe_avg_price_3000', '0_17_all', 'cafe_count_2000_price_500', 'work_all', 'cafe_sum_3000_max_price_avg', '16_29_female', '7_14_all', 'big_church_count_5000', 'ttk_km', 'young_male', '0_6_all', 'children_preschool', 'cafe_count_3000_price_1500', 'leisure_count_3000', '0_13_male', '7_14_male', 'cafe_count_5000_price_2500', 'ekder_all', 'cafe_sum_1000_min_price_avg', 'work_male', 'metro_min_walk', 'cafe_count_5000_na_price', '0_17_male', 'cafe_sum_500_min_price_avg'

В дальнейшем, эксперт в области недвижимости может в ручную проанализировать, что подлежит удалению, а что следует оставить.

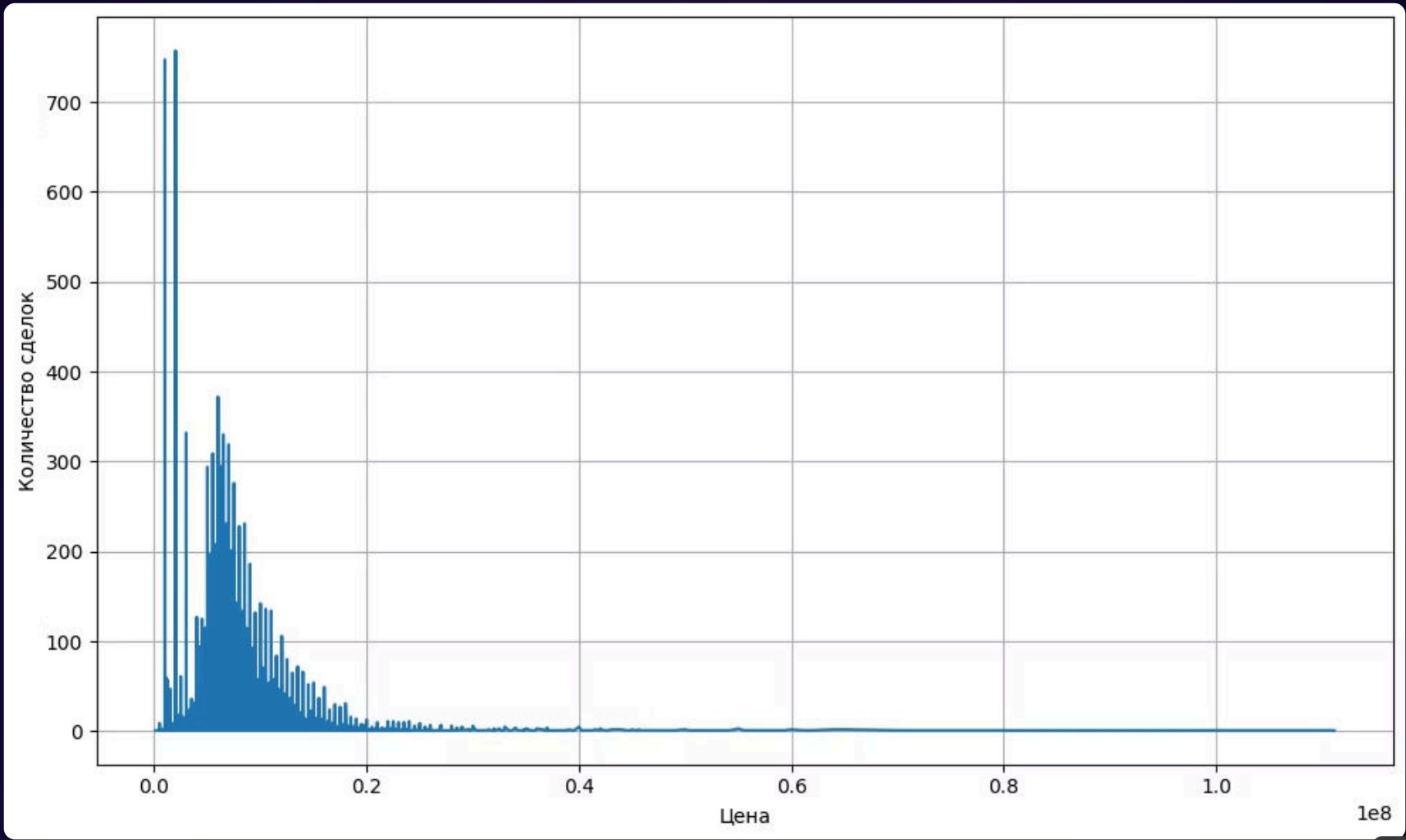
Выявление аномалий

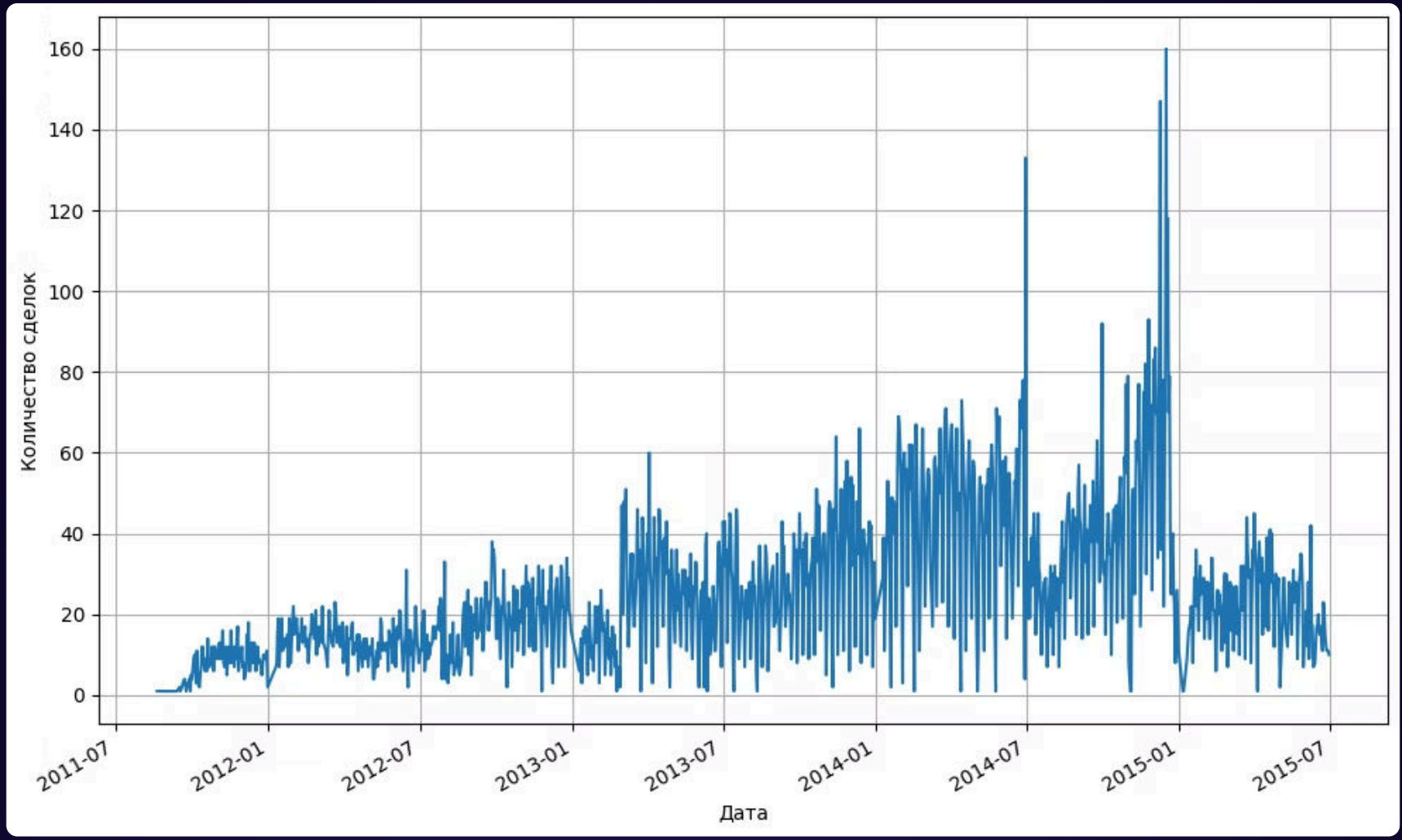
Учитывая количество полей(292), защитимся от всех возможных выбросов и аномалий с помощью фильтрации по квантилю.

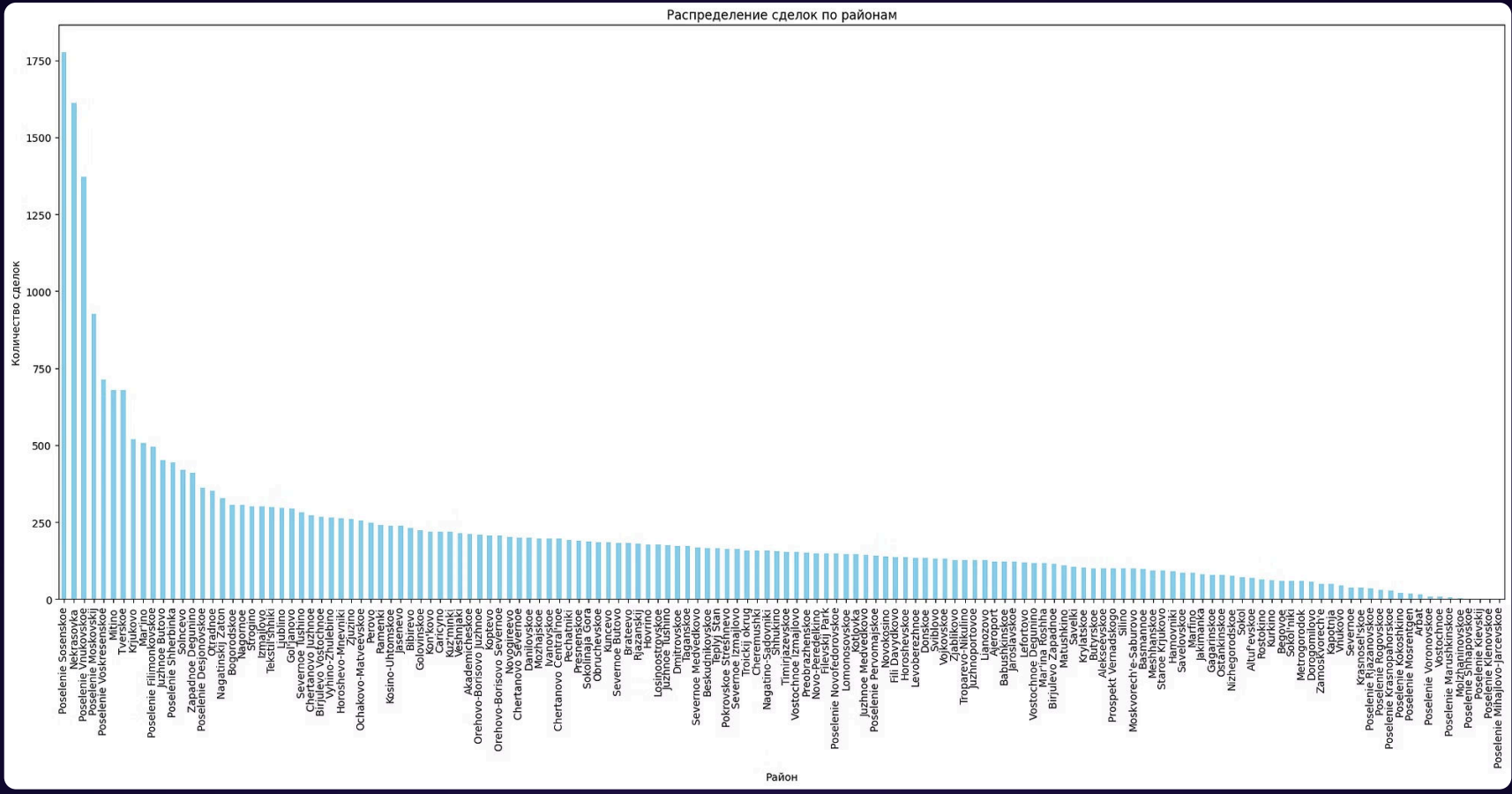
Оставшиеся отклонения(возникшие из за нехватки значения квантиля под определённое поле) удалим вручную, также с помощью фильтрации.

Сбалансированность данных

Проверим баланс по цене, временному периоду и количеству сделок по районам:



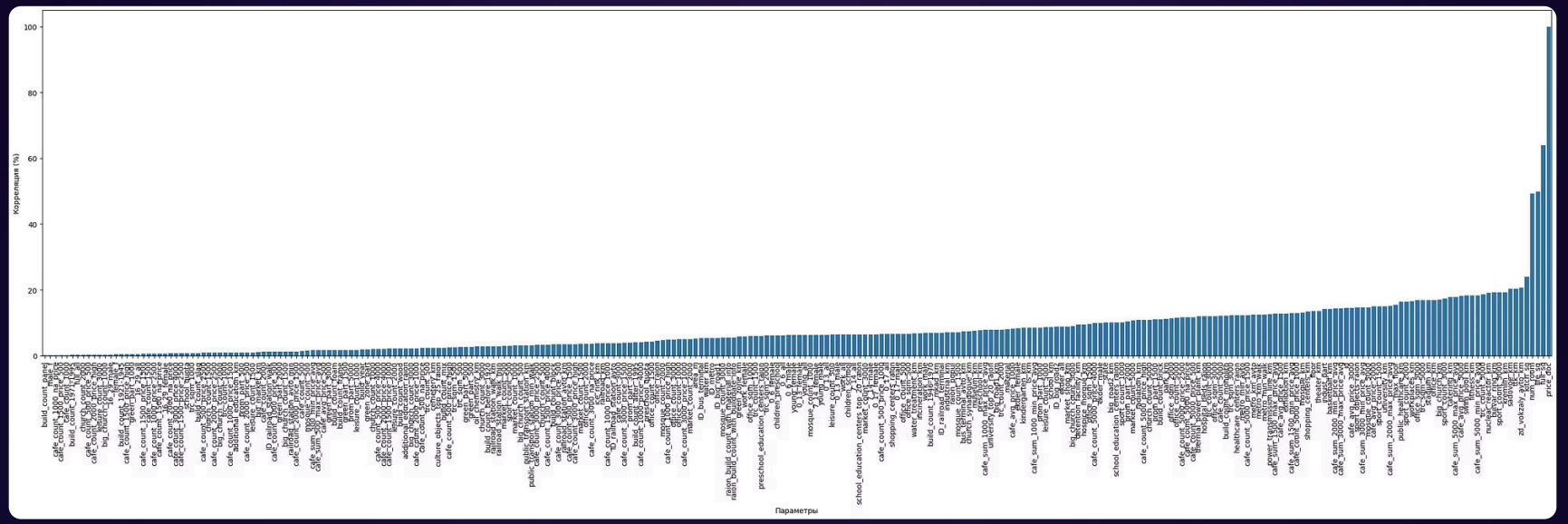




Вывод. Более-менее сбалансированы исходные данные оказались лишь по дате.

Базовый отбор признаков

Чтобы проверить влияние полей на целевую переменную(цену) составим корреляционную матрицу, а затем изобразим полученный результат в виде графика(оригинал доступен в блокноте Jupiter):



Статистики

Посчитаем основную статистику в области недвижимости - среднюю цену за квадратный метр. Современные данные взяты с [Цены на недвижимость в Москве на графике за 10 лет в рублях (irn.ru)](irn.ru).