Анализ данных о качестве воздуха

Выбор данных

OpenAQ Platform

- Почасовые измерения по всему миру
- Указание локации
- Любой временной промежуток
- 5 показателей
- Всего - любое кол-во записей

Air Quality Data Set

- Почасовые измерения по Италии
- Без указания локации
- За 390 дней
- 13 показателей
- Всего – 9471 записей

Выбор данных

- Исходя из задачи, а именно, требований проанализировать данные за промежуток 3х лет, а так же в нескольких городах, были выбраны данные с OpenAQ Platform.
- Порядок создания датасета:
  - Для дальнейшего использования был написан скрипт их загрузки (parse_data.py)
  - Я определил диапазон дат для загрузки как (с 01.01.2020 по 01.01.2023):
    - DATE_FROM = datetime(2020, 1, 1)
    - DATE_TO = datetime(2023, 1, 1)
  - Далее был загружен список всех стран, которые могут потенциально подойти под критерии
    - Есть данные за указанный промежуток времени
    - Есть все 5 показателей, которые предоставляет платформа
  - После чего был сформирован список всех локаций из этих стран, так же удовлетворяющих всем критериям, описанным выше
  - В итоге был получен датасет с размером 1187785 rows × 13 columns (файл openaq_measurements_2.zip)

Предварительная обработка данных

- В данных были обнаружены измерения без указания локаций, в количестве **157227**, что составило **13.23%** от общего числа.
- Поскольку такие данные трудно будет связать с природными или антропогенными факторами, они были удалены.
- Далее были удалены пропуски и данные были сгруппированы по городам, сами измерения стали средними за час

- После чего я транспонировал по столбцу с параметром, чтобы для каждой записи получить строку вида:
- `city dateUTC latitude longitude unit co no2 pm10 pm25 so2`
- В итоге получился размер датасета `377985` rows × `10` columns, который содержал информацию о городе, параметре, значении измерения, дате и времени, широте и долготе измерения, единицах измерения.

Предварительная обработка данных

Исходя из описания датасета можно заметить, что наблюдается какое-то количество отрицательных значений.

Данные показания сенсоров можно считать невалидными, т.к. на ресурсе не было указано, что они могут быть относительными, значит, будем считать, что это прямые измерения. Соответственно, прямые измерения концентрации веществ в воздухе должны быть строго положительными.

print(pivot_data.shape)

Было: (377985, 10)

Стало: (341963, 10)

Потеряно: 377985 - 341963 = **36022**

parameter	latitude	longitude	co	no2	pm10	pm25	so2
count	377985.000000	377985.000000	377985.000000	377985.000000	377985.000000	377985.000000	377985.000000
mean	43.255981	-12.733698	9.303396	-16.256183	8.768041	3.329338	-69.378983
std	9.899970	45.455159	372.653589	143.946280	36.076843	36.950989	256.386065
min	20.578611	-121.940560	-4600.000000	-999.000000	-999.000000	-999.000000	-999.000000
25%	36.067050	1.539138	0.000000	0.000000	0.000000	0.000000	0.000000
50%	49.156110	4.923300	0.000000	0.000000	0.000000	0.000000	0.000000
75%	49.747330	14.197074	0.000000	1.560000	12.600000	5.000000	1.000000
max	54.319230	18.089306	174000.000000	381.640000	3559.500000	1156.000000	1541.000000

Предварительная обработка данных

Теперь нужно убрать все города, у которых много пропусков.

Построим топ 10 городов по количеству дней с измерениями (почасовые измерения за день были усреднены и считались как 1 измерение за день).

Отберем только европейские страны (соответственно города) для того, чтобы измерения были с одного и того же континента.

Получим следующий датасет:

- Почасовые измерения по 6 городам: Мальта, Чехия, Люксембург, Австрия, Андорра.
- С указанием точного места
- За ~1000 дней
- 5 показателей
- Без пропусков и невалидных значений
- Всего – 151215 записей

id	city	date_count	country
0	BRITISH COLUMBIA	1065	CA
1	Plzeň	1058	CZ
2	Studénka	1056	CZ
3	Praha	1049	CZ
4	Msida	1038	MT
5	Escaldes-Engordany	1032	AD
6	Salamanca	1027	MX
7	Air Quality Luxembourg	1014	LU
8	Gharb	1013	MT
9	8020 Graz	969	AT

EDA

Построим матрицу корреляций

1. Latitude (широта):

- Наибольшая положительная корреляция с уровнем CO (0.18) и NO2 (0.17), что может говорить о том, что на более высоких широтах концентрации этих загрязнителей могут быть выше.

- Умеренная отрицательная корреляция с PM10 (-0.31) и PM2.5 (-0.25), что может указывать на то, что в более северных регионах концентрация твердых частиц ниже.

2. Longitude (долгота):

- Умеренная положительная корреляция с CO (0.32) и NO2 (0.26), что может говорить о зависимости концентрации этих загрязнителей от географического положения на восток или запад.

- Отрицательная корреляция с SO2 (-0.21), что предполагает снижение уровня этого загрязнителя в зависимости от долготы.

3. CO и NO2:

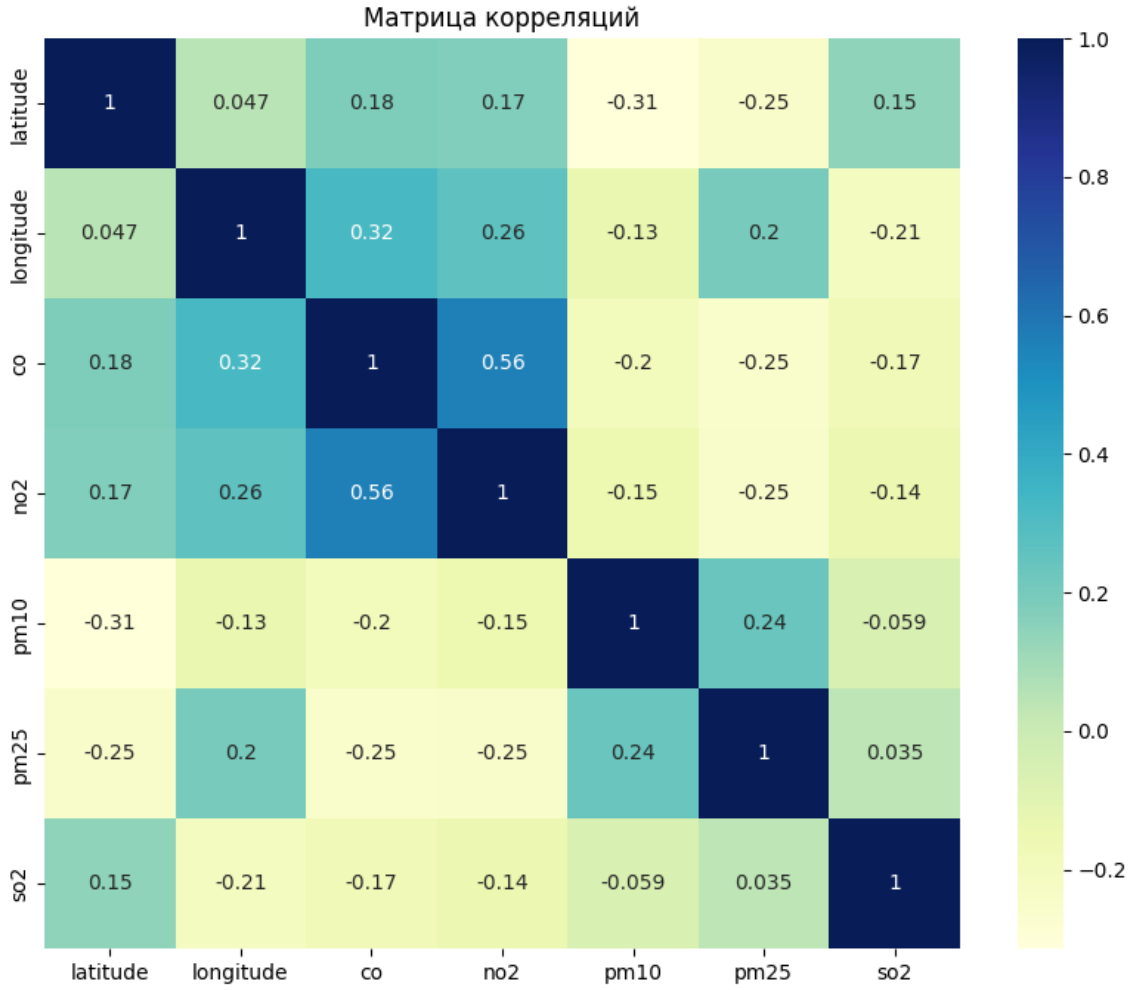
- Высокая положительная корреляция между CO и NO2 (0.56) говорит о том, что эти загрязнители часто появляются вместе, возможно, из схожих источников).

4. PM10 и PM2.5:

- Между PM10 и PM2.5 наблюдается положительная корреляция (0.24), что логично, так как оба показателя отражают концентрации твердых частиц в воздухе.

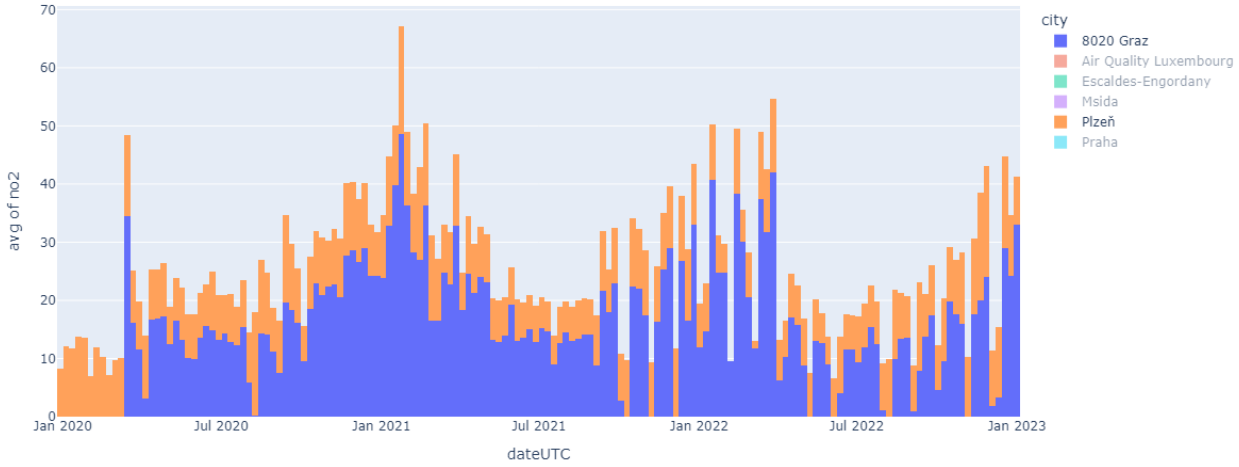
5. SO2:

- SO2 не имеет сильных корреляций с другими загрязнителями, что может указывать на его независимое распределение в атмосфере, связанное с различными источниками выбросов (например, промышленность).

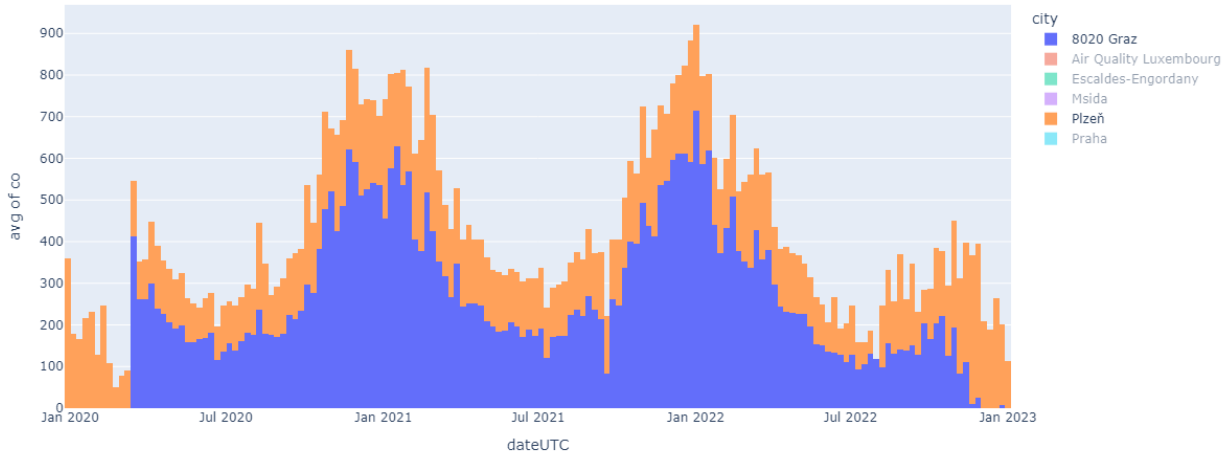


EDA

Значение концентрации NO2 по городу City и дате



Значение концентрации CO по городу City и дате

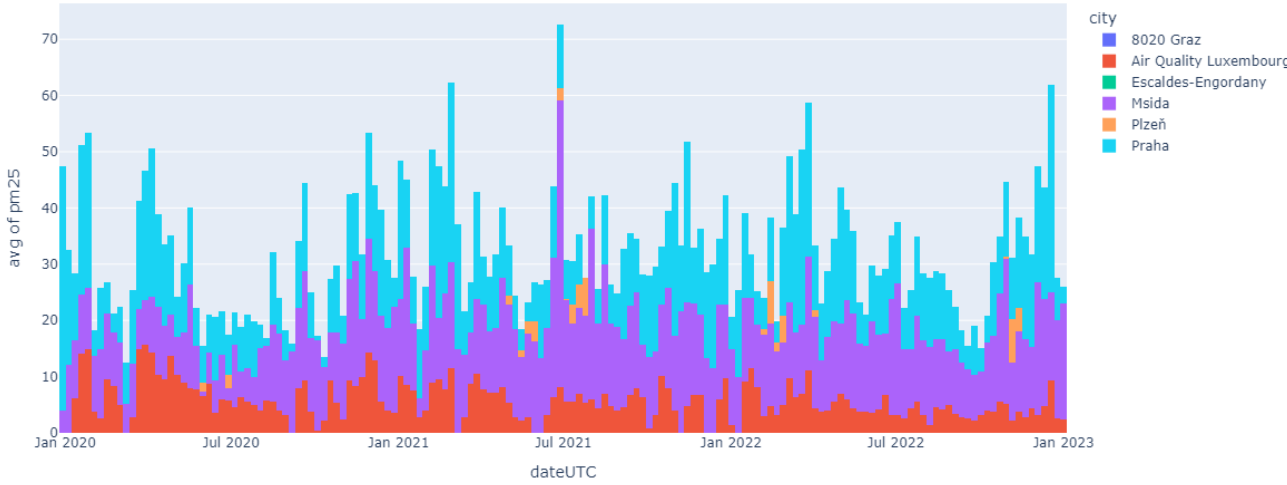


Исходя из полученных визуализаций можно сделать вывод о сезонности измерений CO и NO2

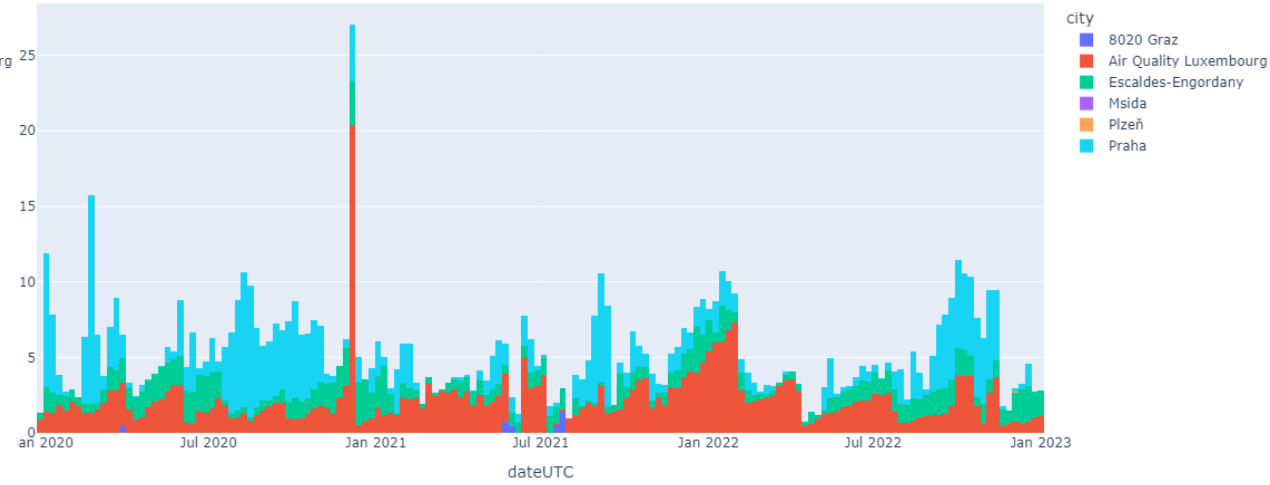
*На графиках представлены только 2 города, т.к. остальные данные неполные и были очищены (пустые графики)

EDA

Значение концентрации pm2.5 по городу и дате



Значение концентрации SO2 по городу и дате



Стандартизация (z-преобразование)

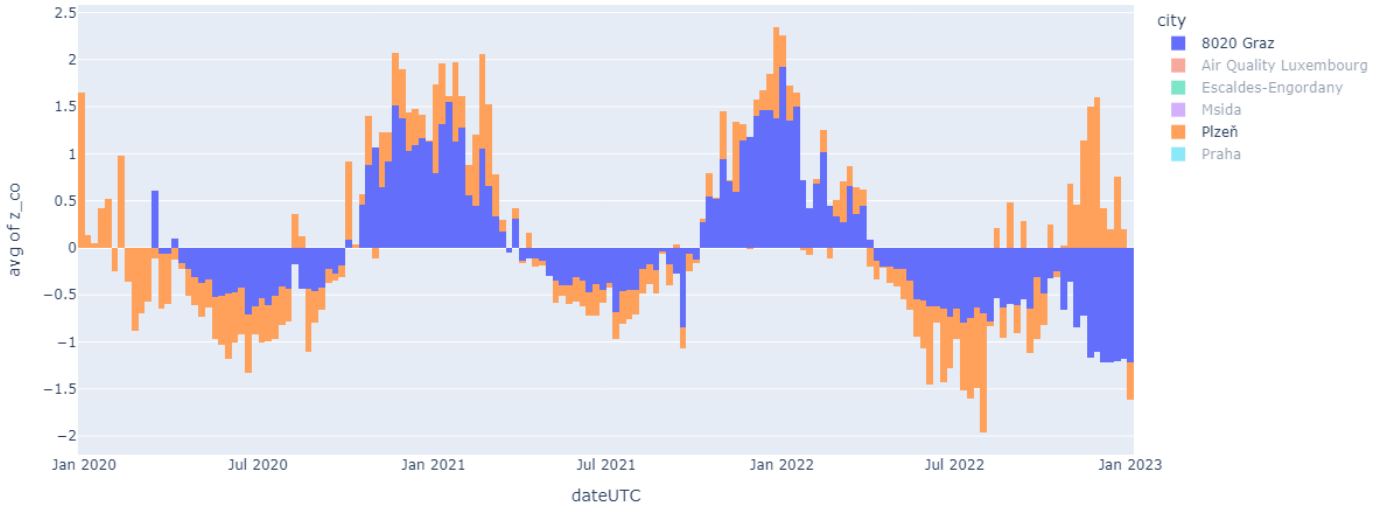
- Данный метод был выбран потому что:
  - Z-преобразование позволяет привести все переменные к единому масштабу
  - Позволяет легко выделить выбросы
  - Используя z-преобразование, можно сравнивать концентрации разных загрязнителей, несмотря на их разный масштаб

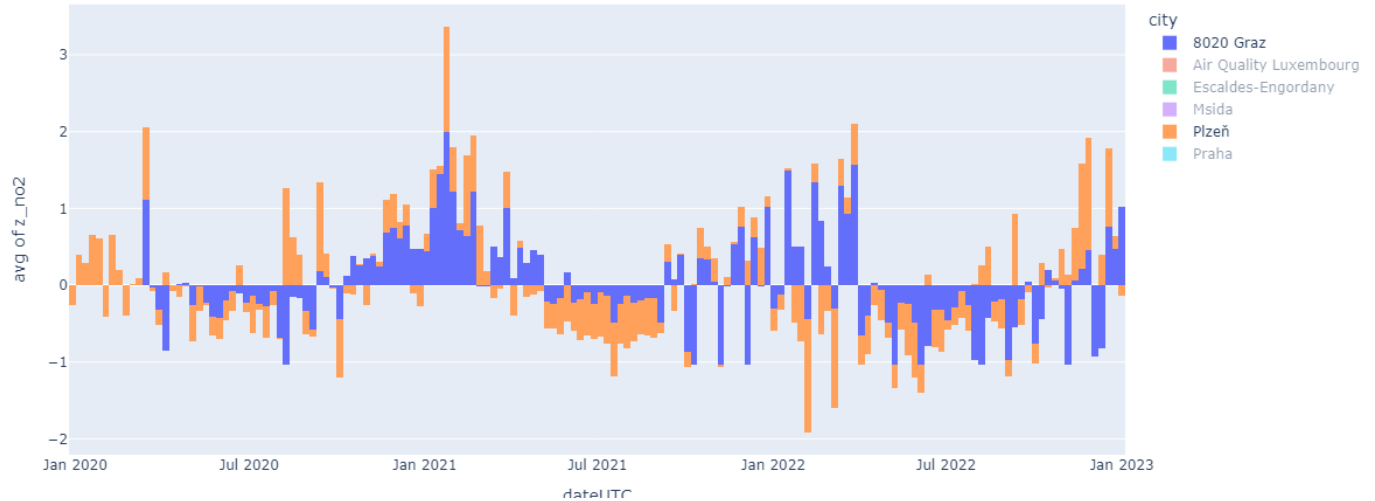
Стандартизация (z-преобразование)

После преобразования еще лучше видна сезонность данных по NO2 и CO2.

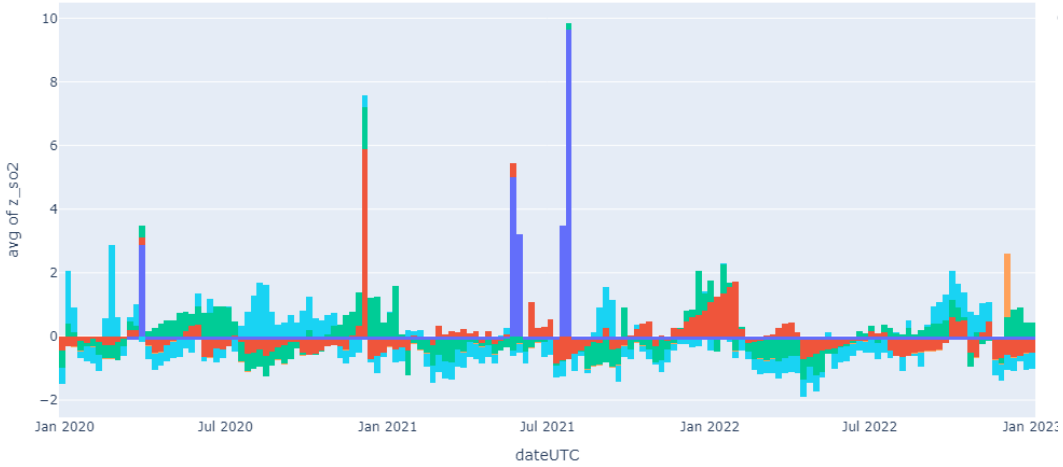
Так можно сразу отсечь аномалии, значения у которых > 3.

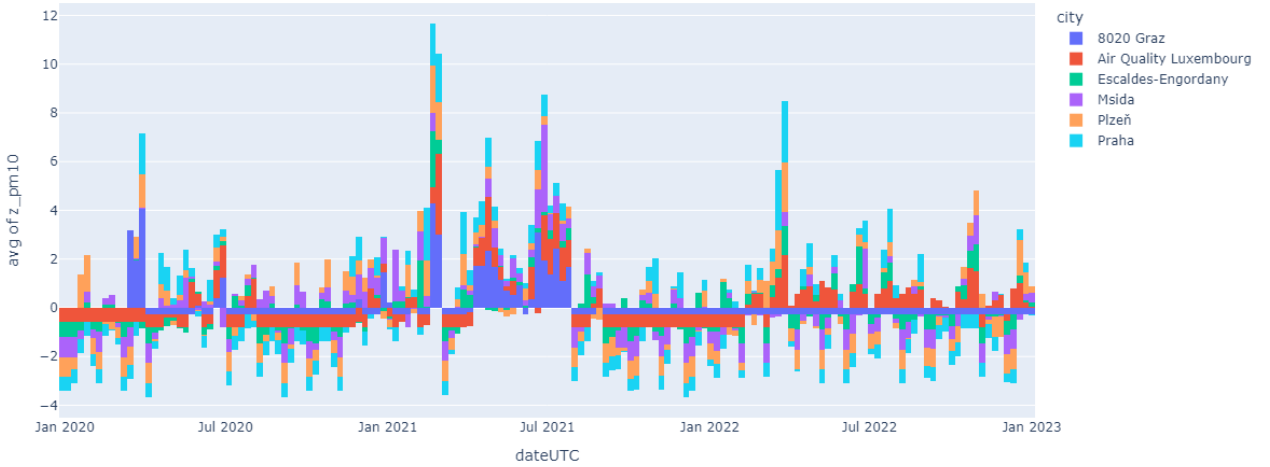
Аномалии наблюдаются по no2 зимой 2021.

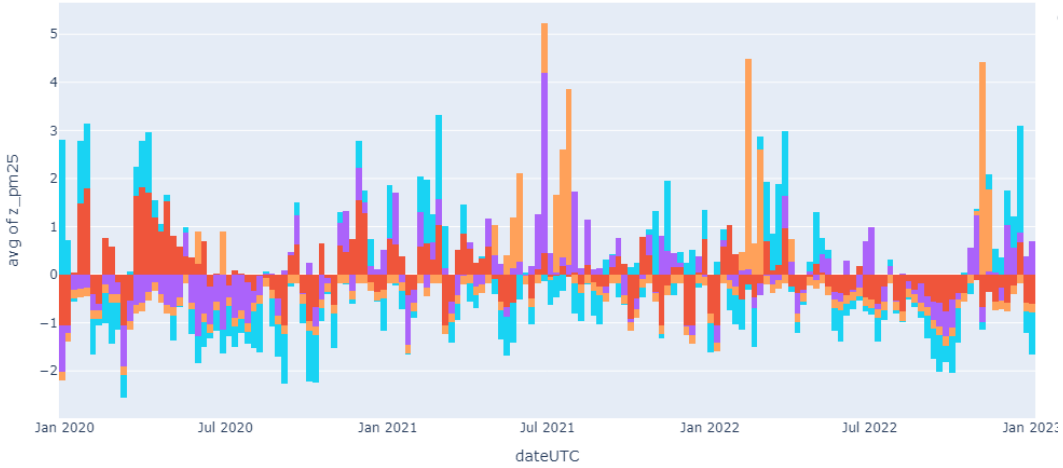




Стандартизация (z-преобразование)







По данным SO2, pm10 и pm2.5 явно выраженной сезонности по временам года нет.

Аномалии наблюдаются по всем, но так же без конкретной привязке к локации или времени года

Самые значительные аномалии наблюдаются в Мальте и в Пльзене по значениям pm25 и pm10, а так же в Праге по pm10 в период зимы-весны 2021 года.

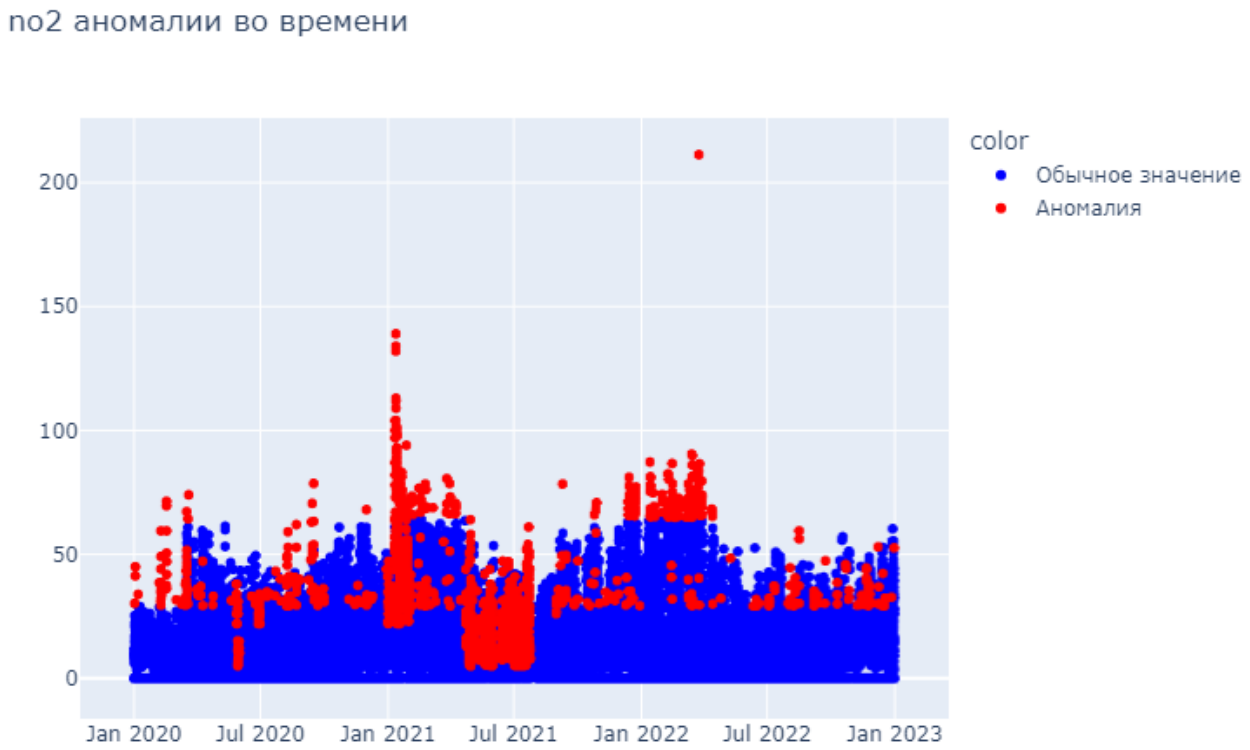
Можно сделать вывод, что данные параметры больше зависят от антропогенных факторов конкретной локации.

Стандартизация (z-преобразование)

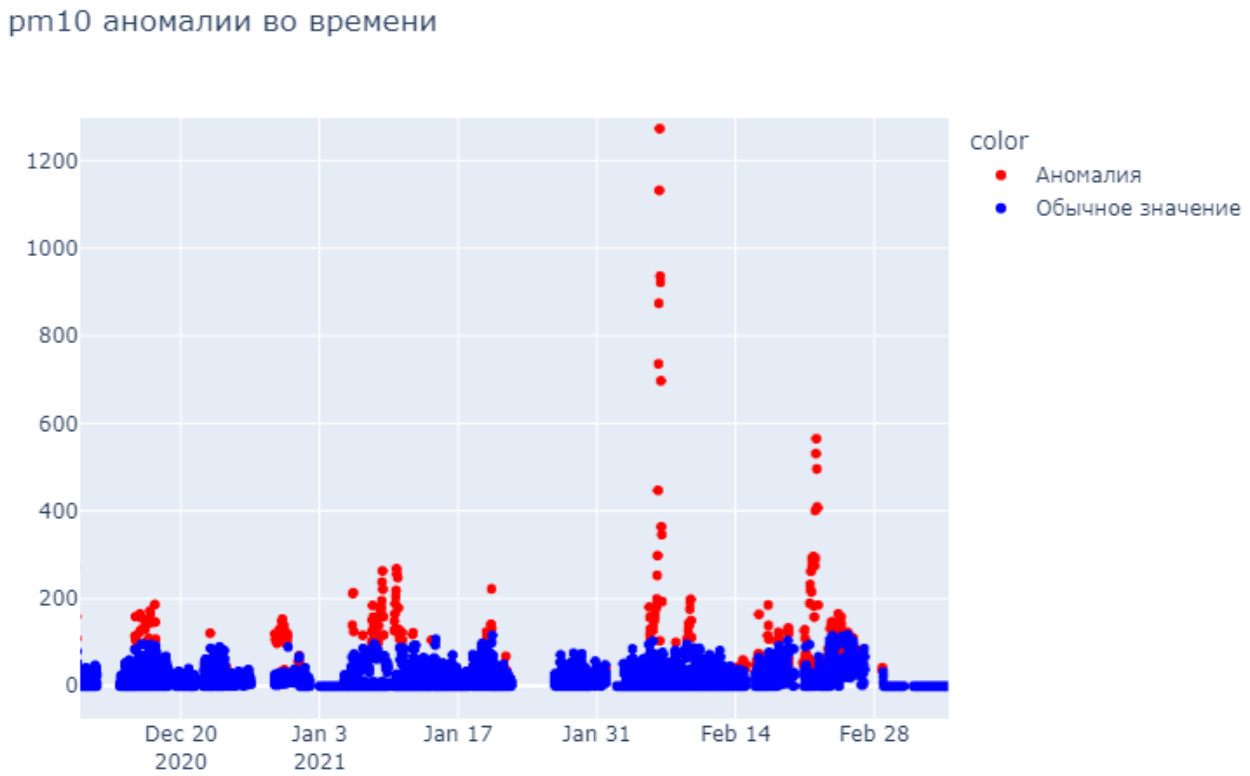
Локация	Аномальных значений для co	Аномальных значений для no2	Аномальных значений для pm10	Аномальных значений для pm25	Аномальных значений для so2	Всего измерений	% аномалий от общего числа
8020 Graz	358	189	364	0	249	18607	6.23
Air Quality Luxembourg	173	415	256	303	31	20408	5.77
Escaldes-Engordany	92	408	232	0	256	20860	4.74
Msida	624	0	395	388	0	21440	6.56
Plzeň	115	342	384	459	5	22159	5.89
Praha	249	351	414	304	138	21086	6.90

Стандартизация (z-преобразование)

Стандартизация (z-преобразование)



Стандартизация (z-преобразование)



Стандартизация (z-преобразование)

Стандартизация (z-преобразование)

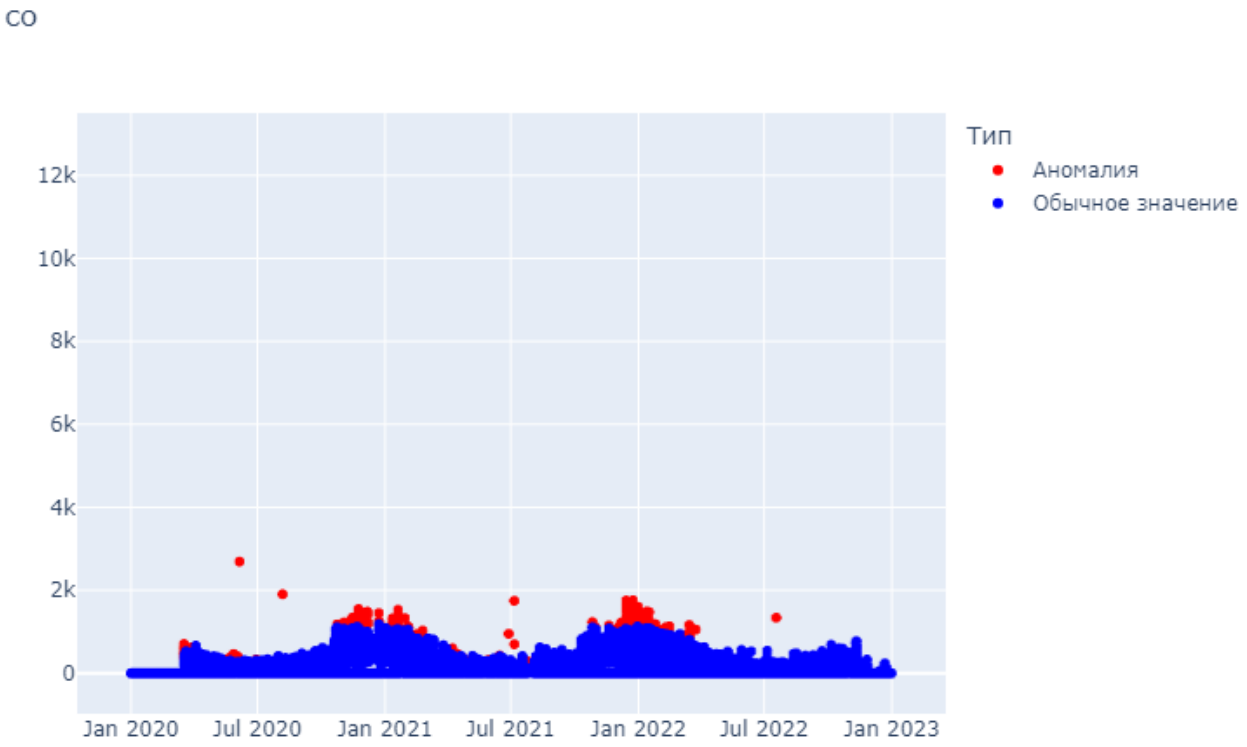
Дополнительные методы поиска аномалий

- Изолирующий лес (Isolation Forest)
- Локальная оценка выбросов (LOF)
- Метод соседей (k-nearest neighbors)
  - Не был выбран потому что требует заранее известного числа кластеров, что не всегда возможно в экологических данных.

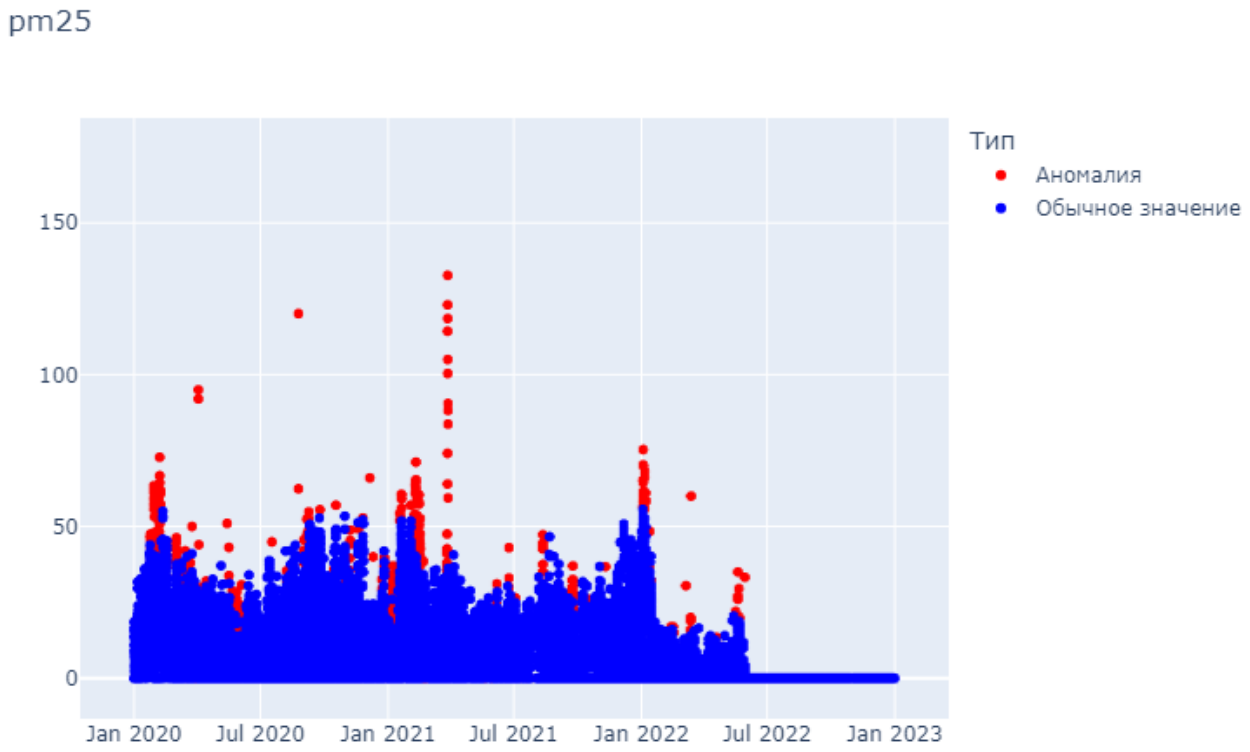
Изолирующий лес (Isolation Forest)

- Данный метод был выбран потому что:
  - Isolation Forest хорошо справляется с задачей обнаружения аномалий в данных с большим количеством признаков
  - Не требует предположений о нормальном распределении данных
  - Обладает хорошей производительностью на больших наборах данных

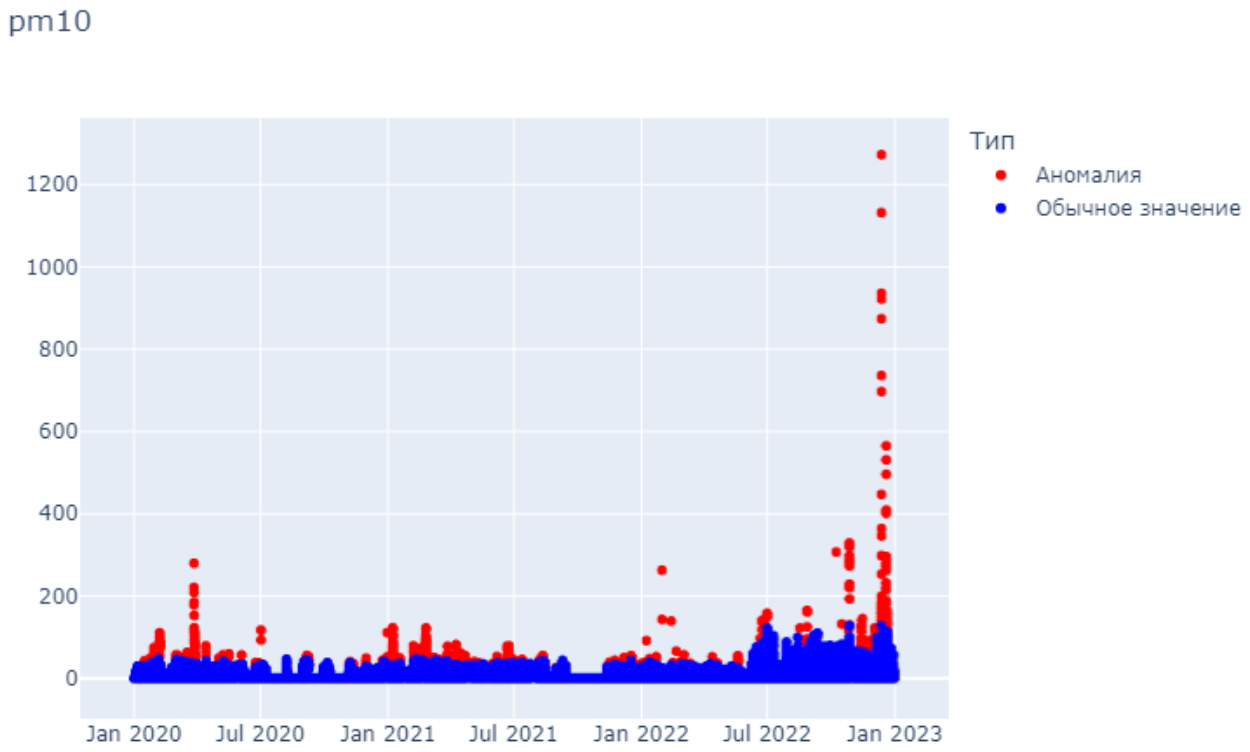
Изолирующий лес (Isolation Forest)



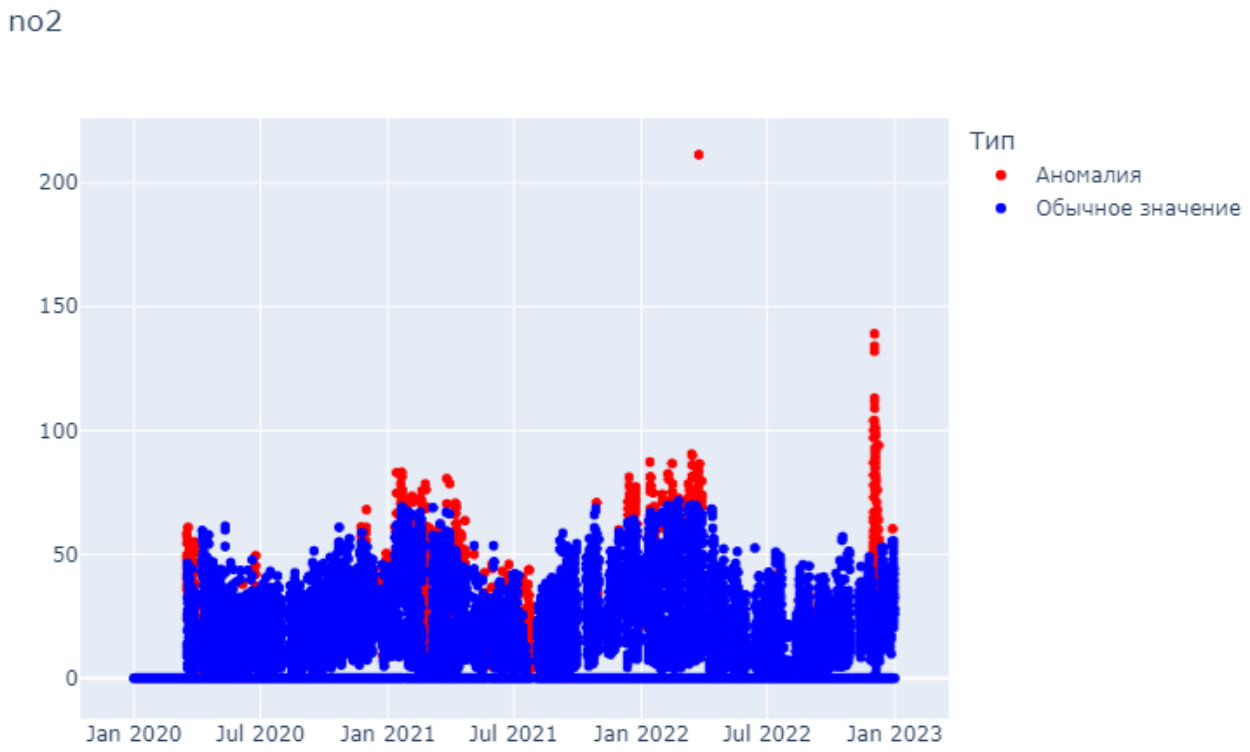
Изолирующий лес (Isolation Forest)



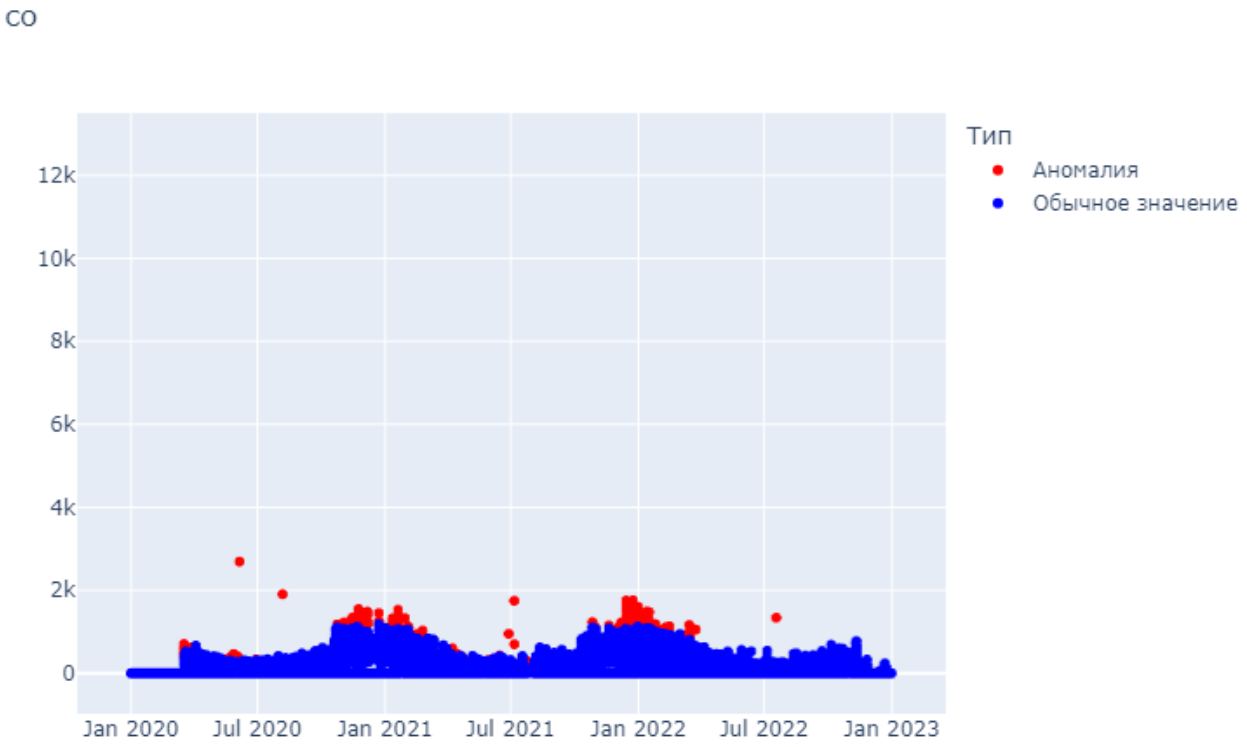
Изолирующий лес (Isolation Forest)



Изолирующий лес (Isolation Forest)



Изолирующий лес (Isolation Forest)



Изолирующий лес (Isolation Forest)

Полученные результаты похожи на те, что были представлены ранее с помощью стандартизации.

Данный метод так же не показал сезонной зависимости по аномалиям

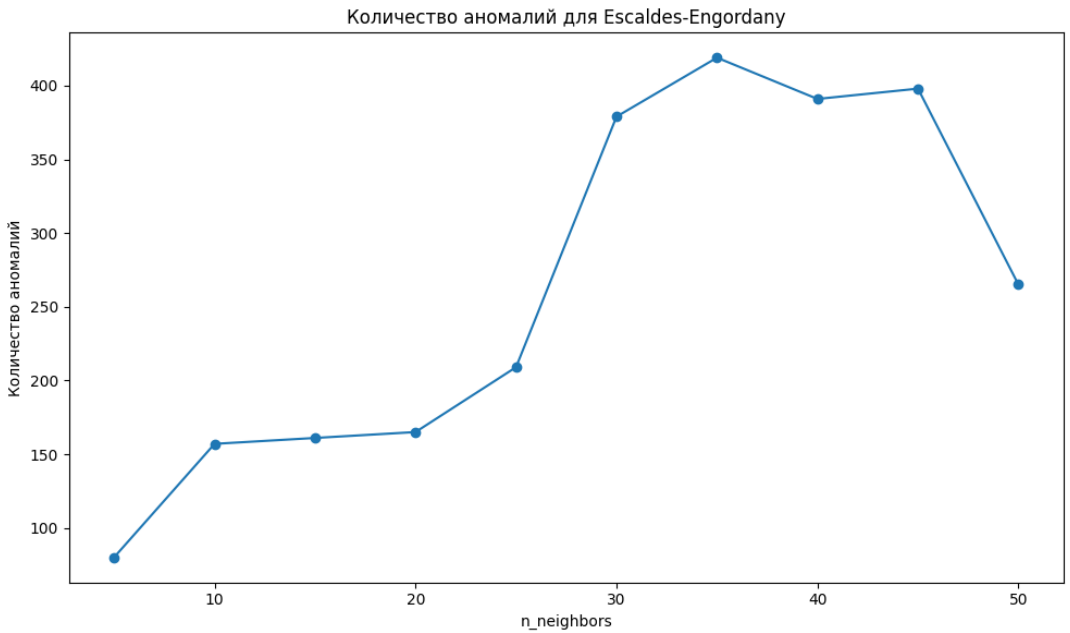
Общий процент аномалий: 5.00%

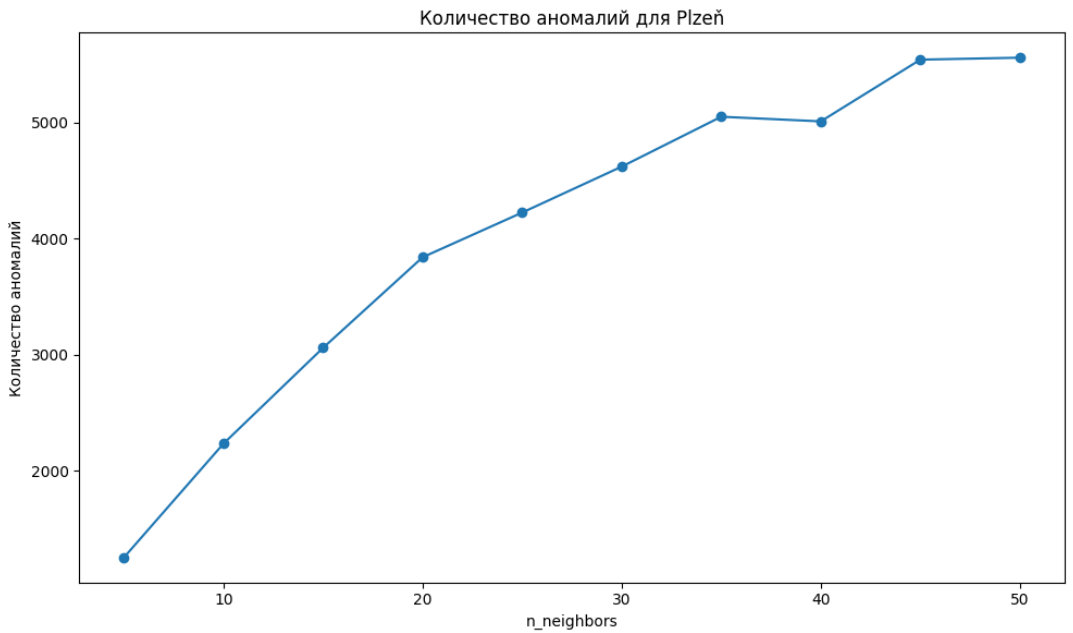
Локация	% аномалий
8020 Graz	5.003493
Air Quality Luxembourg	4.998040
Escaldes-Engordany	5.000000
Msida	4.995336
Plzeň	5.000226
Praha	5.003320

Локальная оценка выбросов (LOF)

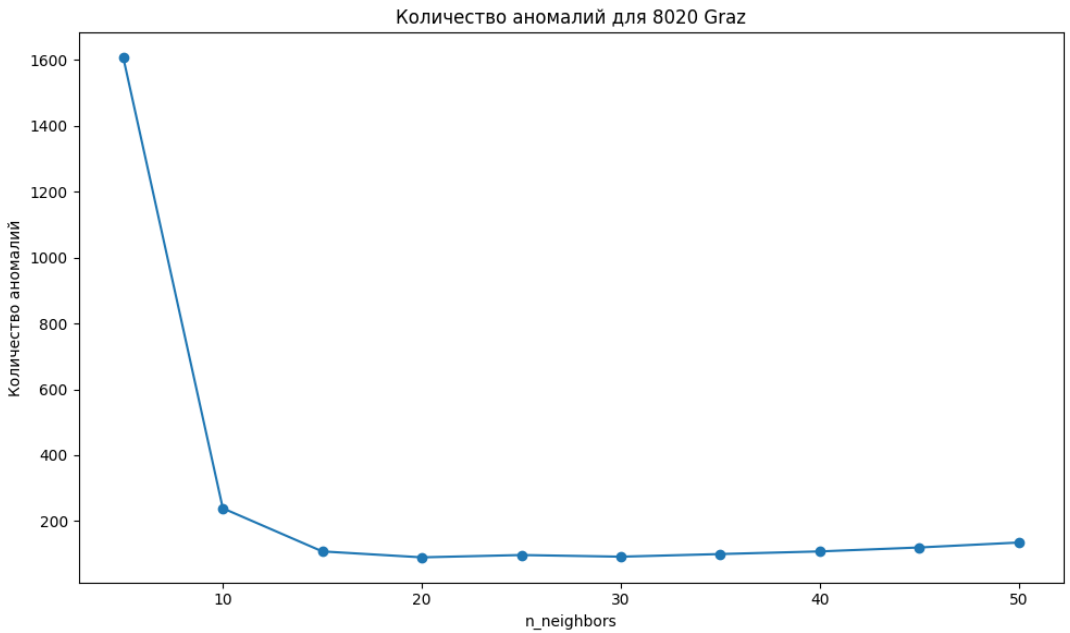
- Данный метод был выбран потому что:
  - LOF оценивает аномалии, сравнивая плотность данных в локальной области с плотностью соседей. Это важно в случае, если аномалии имеют локальный характер, например, резкие изменения концентраций загрязнителей в определенных городах или периодах времени.
  - Хорошо подходит для наборов данных, где плотность распределения может варьироваться в разных областях пространства признаков (в данном случае у нас отличается плотность в зависимости от города и времени года, в некоторых случаях кол-во измерений разное)
  - Легко работает с многомерными признаками. Он может учитывать зависимость между показателями и находить выбросы в сложных пространственных структурах данных.

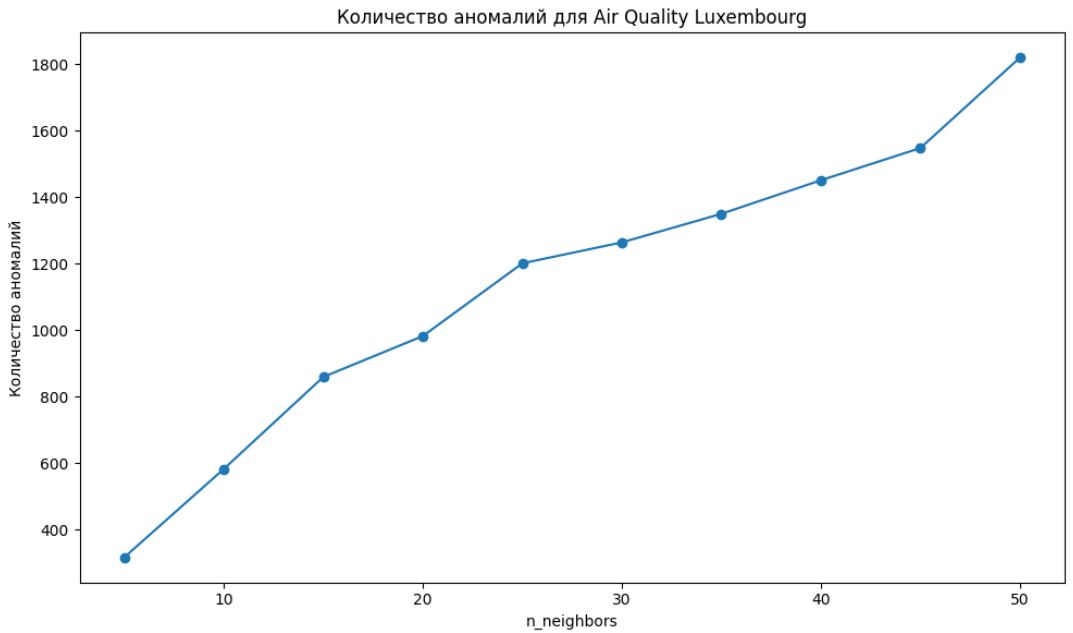
Локальная оценка выбросов (LOF)



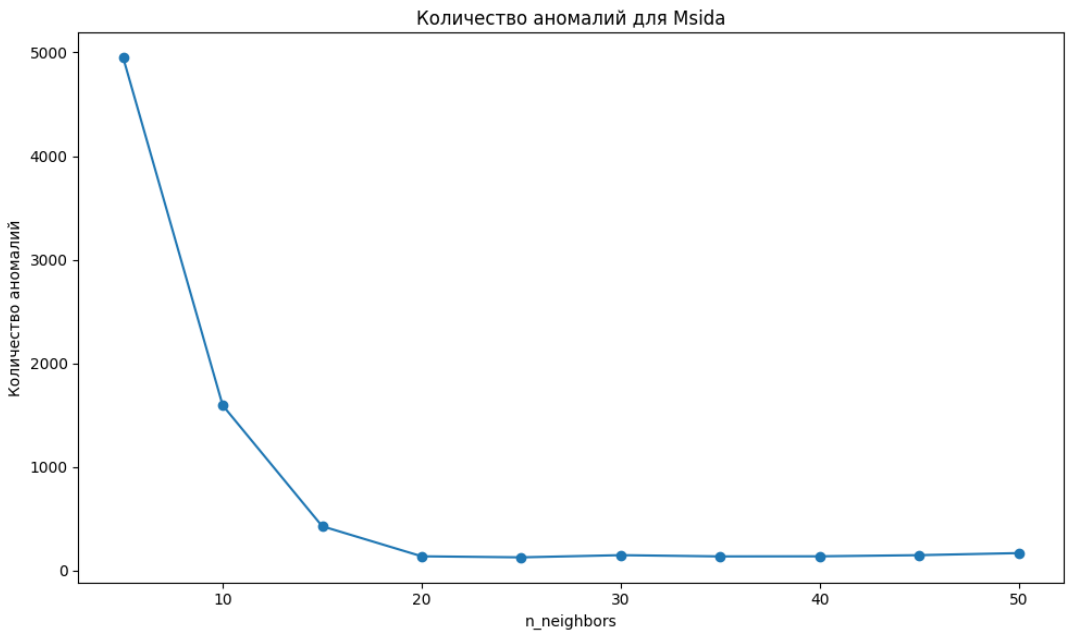


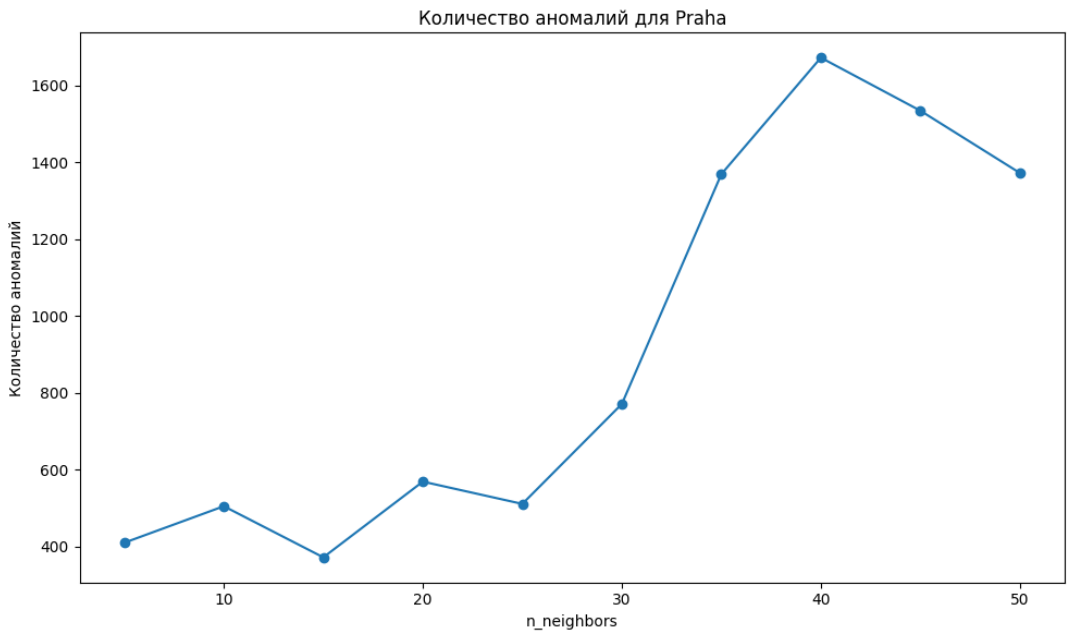
Локальная оценка выбросов (LOF)





Локальная оценка выбросов (LOF)





Локальная оценка выбросов (LOF)

- Исходя из полученных значений можно сказать, что LOF плохо подойдет для дальнейшего использования т.к. на стандартных значениях кол-ва соседей от 5 до 50 он показывает неодинаковые и нестабильное кол-во аномалий
- При дальнейшем увеличении кол-ва соседей продолжается такое же непонятное кол-во аномалий, которое не совпадает с тем количеством, которое было получено предыдущими двумя методами.

Пример для 200 соседей:

Город	Кол-во аномалий
Praha	477
Msida	332
Air Quality Luxembourg	4172
8020 Graz	408
Plzeň	1146
Escaldes-Engordany	1013

Выводы

Причины аномалий:

- Сезонные колебания (есть сезонность – зимой выбросов большем чем летом, это может быть связано с антропогенным фактором - отопление).
- Любые другие антропогенные факторы, например в больших городах наблюдается большее среднее кол-во выбросов, чем в более маленьких городах, поэтому и аномалии в них сильнее.
- Отметить природные факторы труднее т.к. нет данных по метеорологическим параметрам

Предложения по улучшению

- Увеличение частоты и качества измерений: в начале было отмечено, что есть непонятные значения в исходных данных, а так же впоследствии на графиках отсутствовали показания по некоторым из загрязнителей.
- Установление дополнительных показателей: Для более глубокого анализа можно отслеживать другие параметры, такие как влажность, скорость ветра, температура, т.к. корреляция между показаниями концентраций веществ и локацией есть, а более подробной зависимости по данным невозможно.