Загрязнение воздуха в городе Сеул, Южная Корея

Проект подготовили:
Ахтырский Василий,
Бычин Ярослав,
Старшова Дарья

Основные цели проекта

1. Выявление общей тенденции изменения качества воздуха в Сеуле с течением времени. Сравнение качества воздуха с нормативными показателями.
2. Анализ изменений в составе воздуха за определенные промежутки времени:

    i) анализ суточных колебаний изменений концентраций загрязняющих веществ.

    ii) анализ сезонных колебаний.

3. Исследование качества воздуха в районах Сеула. В частности, выявление самого "грязного" района и самого "чистого" района. Построение карты качества воздуха в Сеуле.
4. Анализ корреляций между загрязняющими веществами.
5. Прогнозирование уровня загрязнения.

Первичная обработка файлов

- Measurement_summary

Основные измерения (SO2, NO2, O3, CO, PM10, PM2.5).

- Measurement_info

Информация о измерениях (дата, станция, статус прибора).

- Measurement_item_info.csv

Нормативные показатели загрязняющих веществ.

- Measurement_station_info

Информация о станциях мониторинга.

Особенности при обработке:

0 - в норме

1 - требуется калибровка

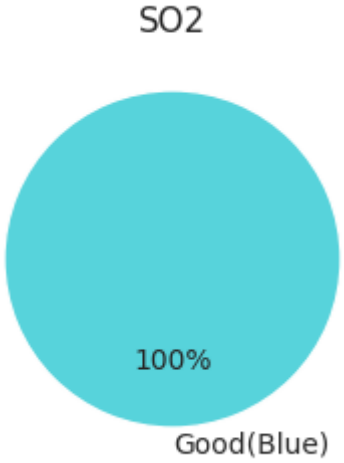
2 - нештатная ситуация

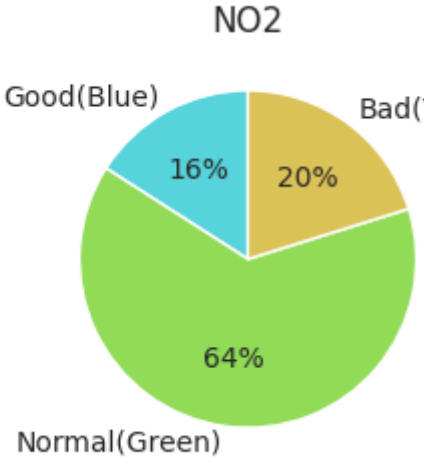
4 - отключение питания

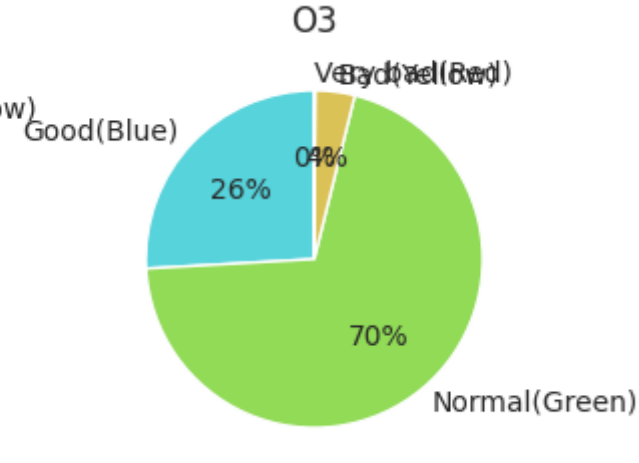
8 - ремонтируется

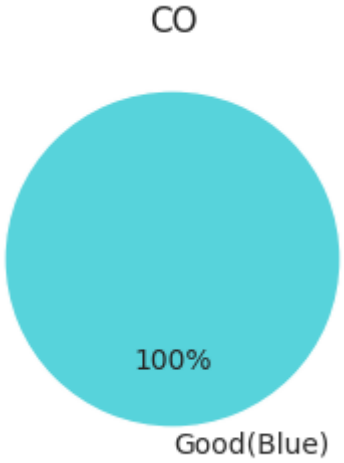
9 - нештатные данные

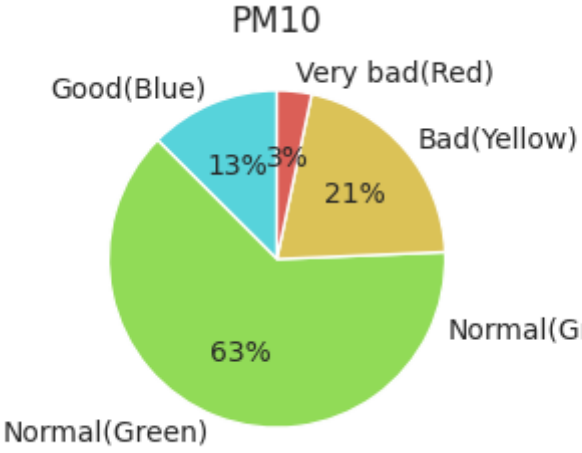
Вывод после анализа файлов: 97% - в норме, 3% - некорректные измерения — удаляем их!

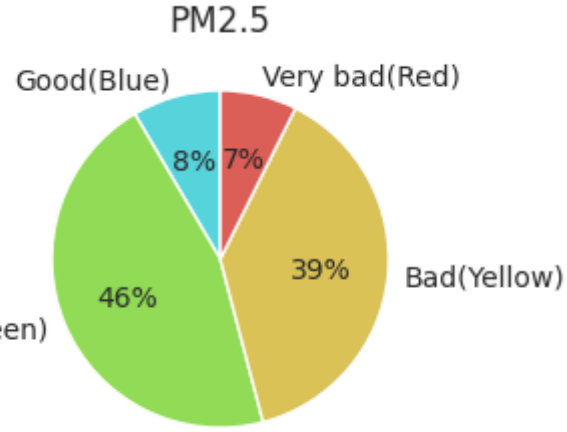

SO2
100%
Good(Blue)

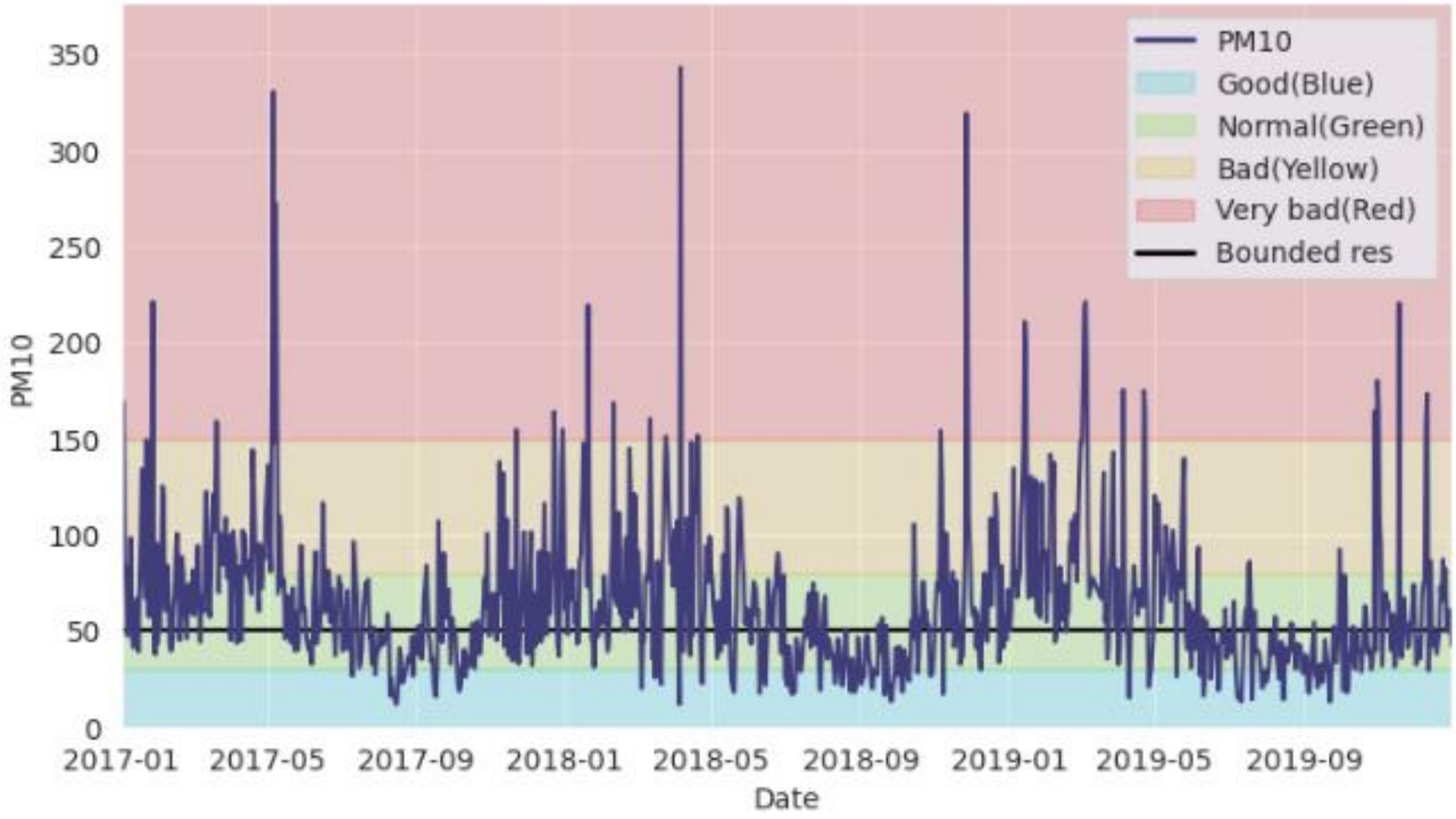

NO2
Good(Blue)
Bad(Yellow)
16%
20%
64%
Normal(Green)

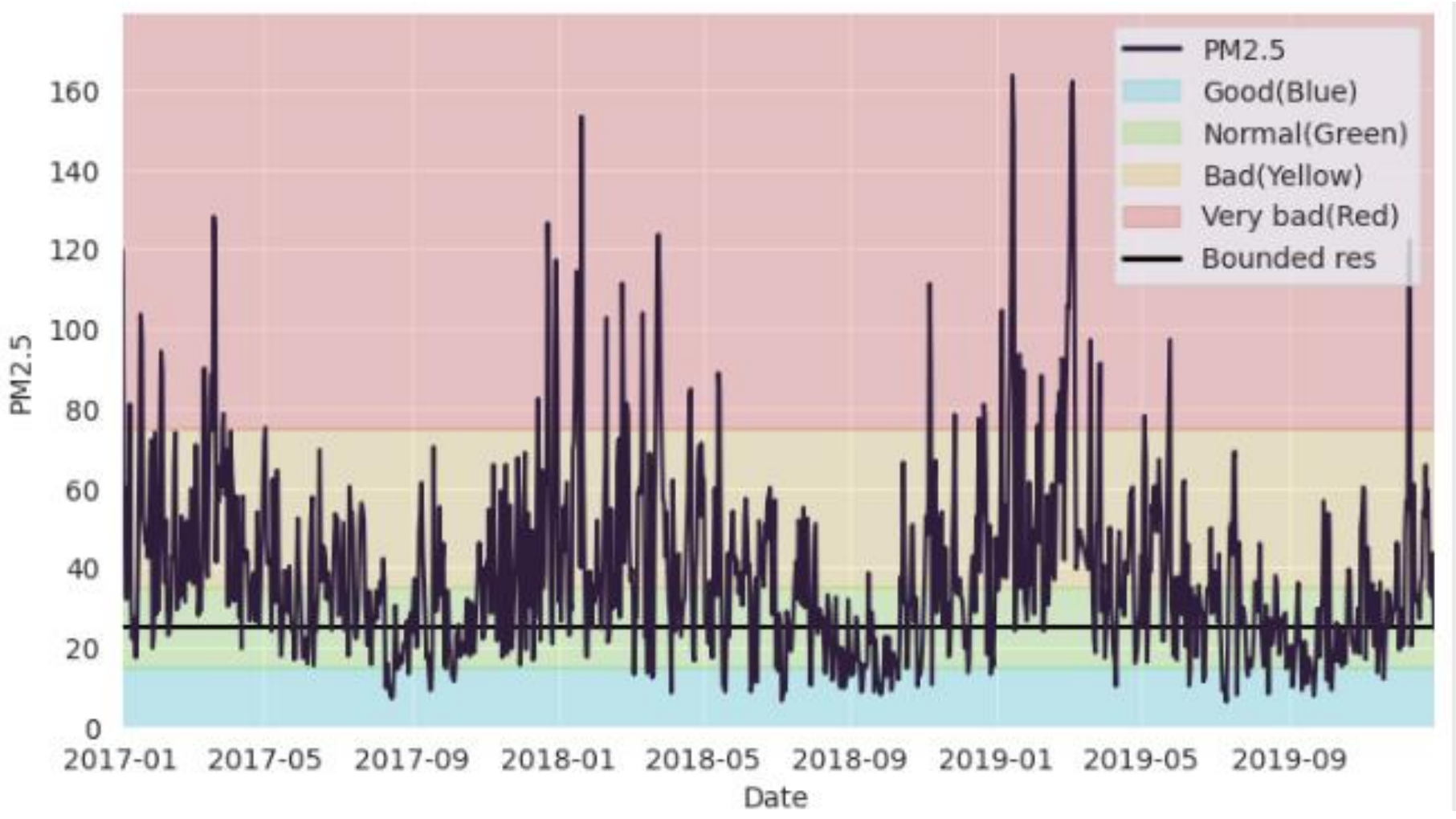

O3
Good(Blue)
26%
4%
70%
Normal(Green)


CO
100%
Good(Blue)

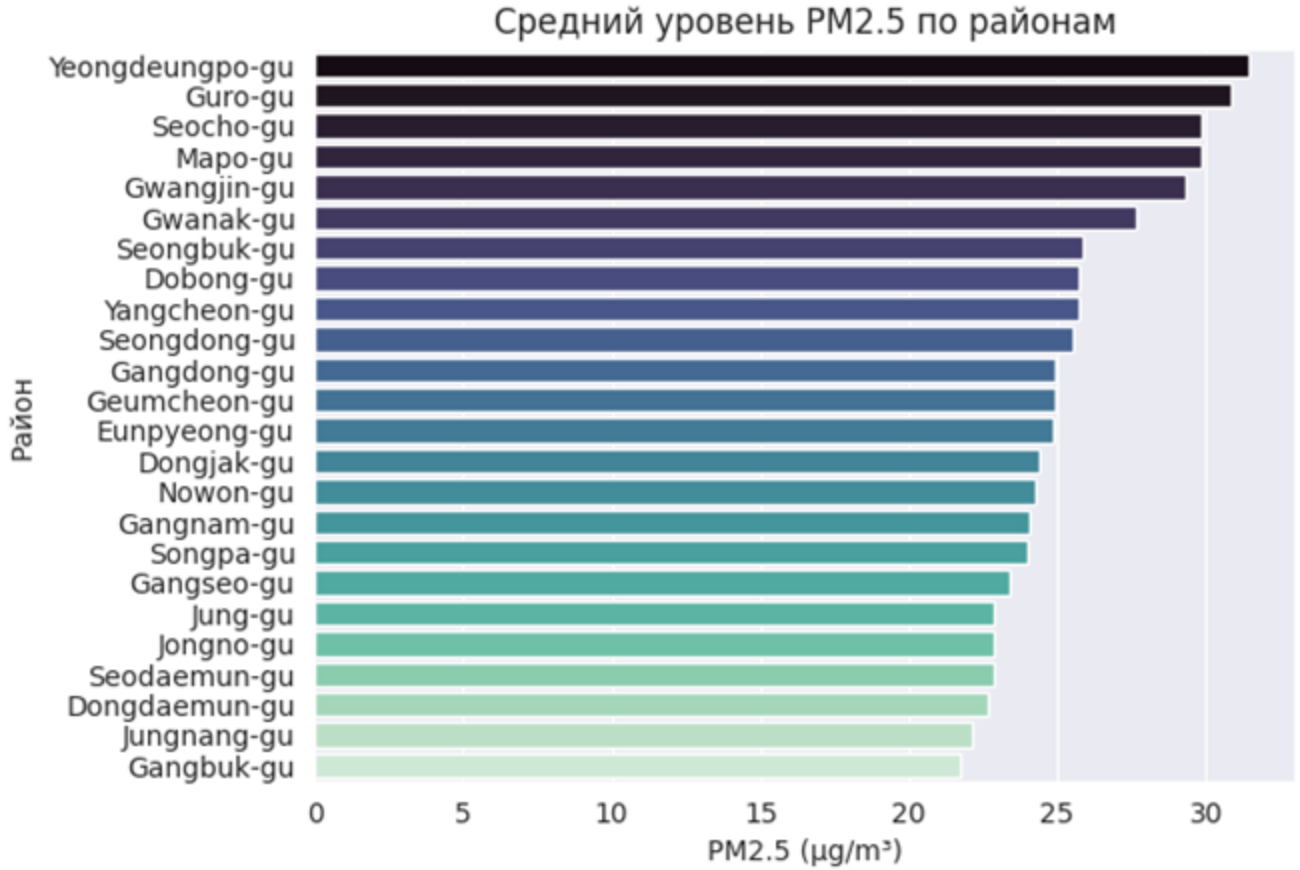

PM10
Good(Blue)
Very bad(Red)
Bad(Yellow)
13%
3%
21%
63%
Normal(Green)

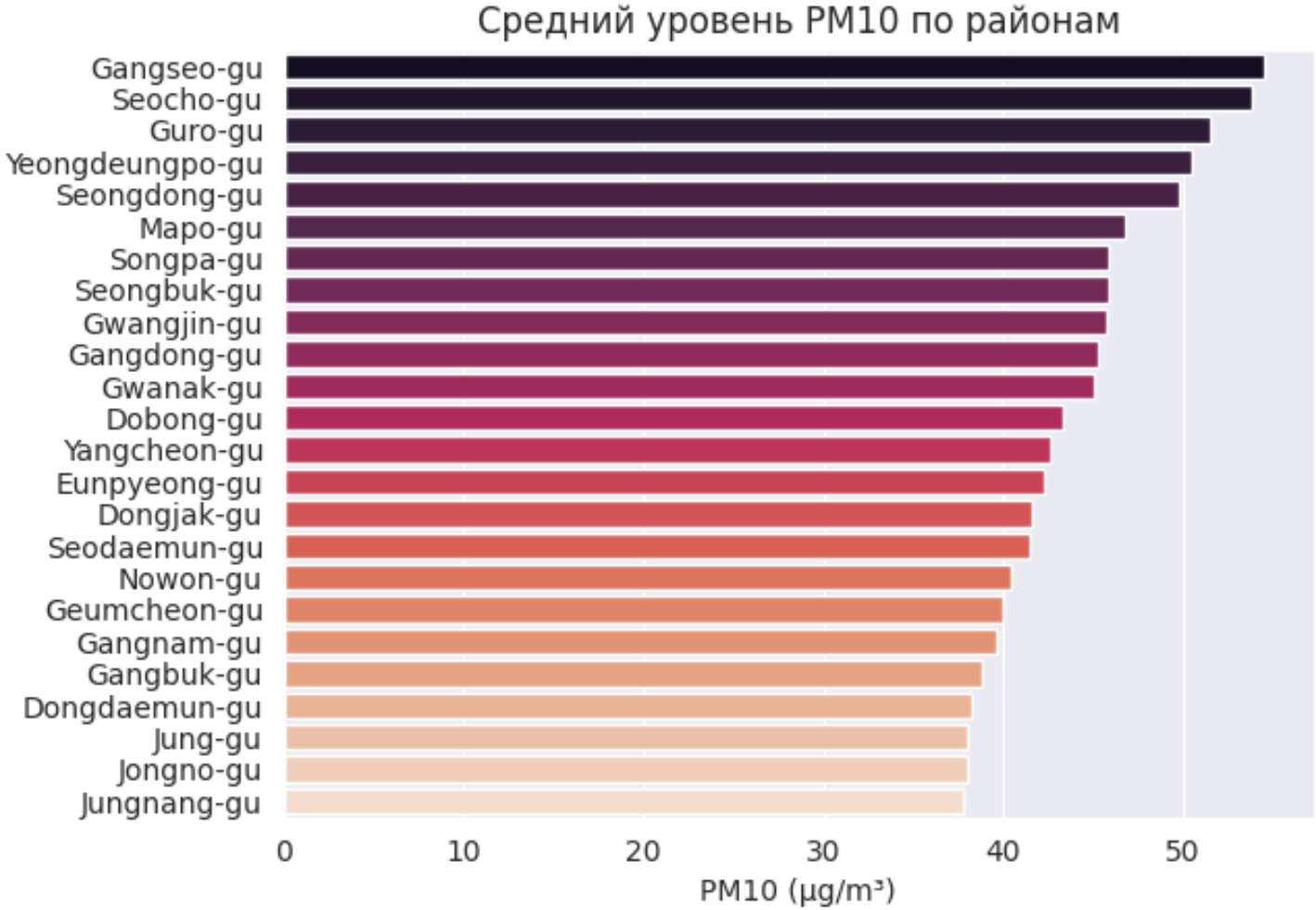

PM2.5
Good(Blue)
Very bad(Red)
8%
7%
39%
Bad(Yellow)
46%
Normal(Green)

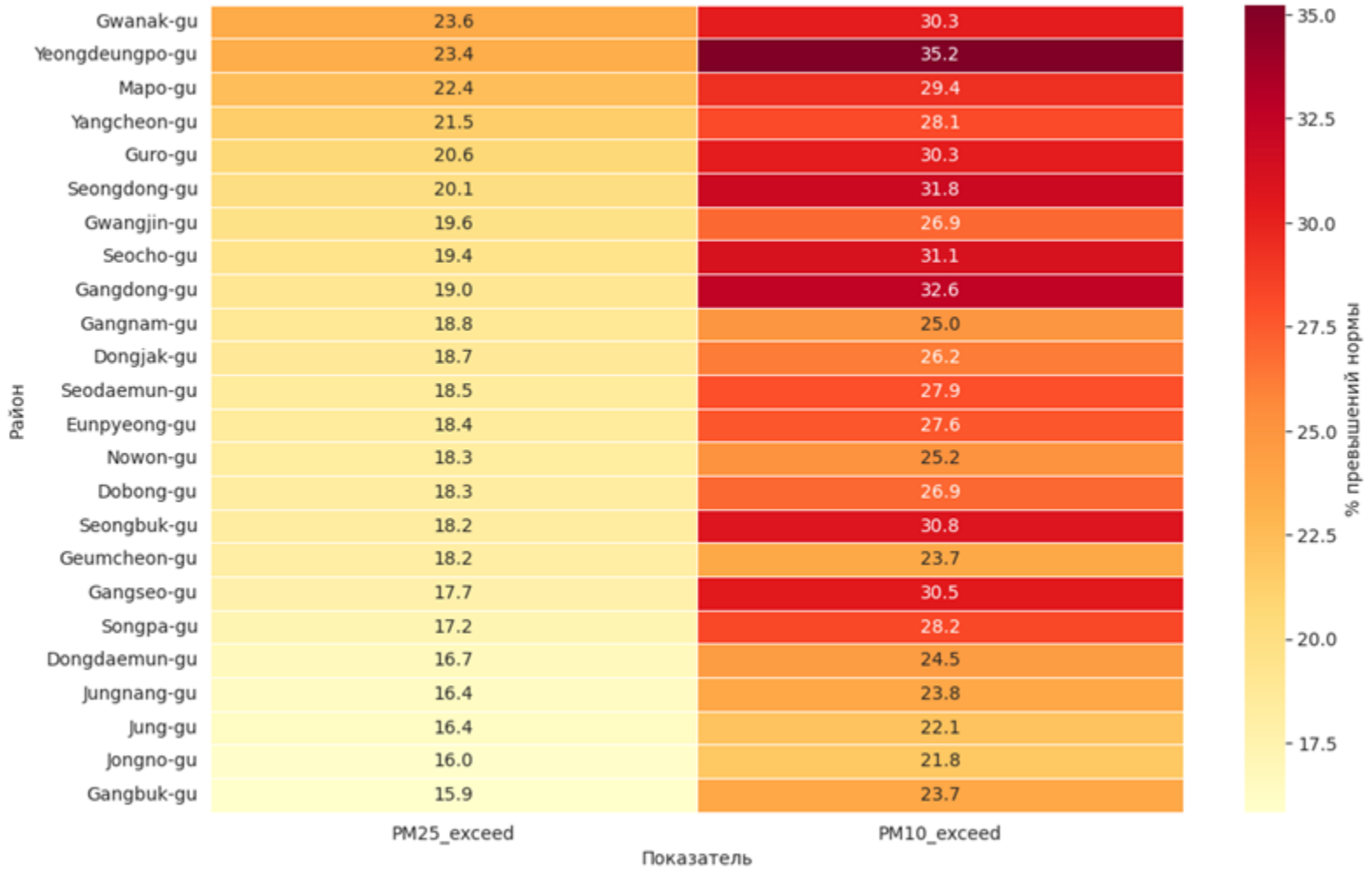

PM10
Good(Blue)
Normal(Green)
Bad(Yellow)
Very bad(Red)
Bounded res
350
300
250
200
150
100
50
0
PM10
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
2019-05
2019-09
Date

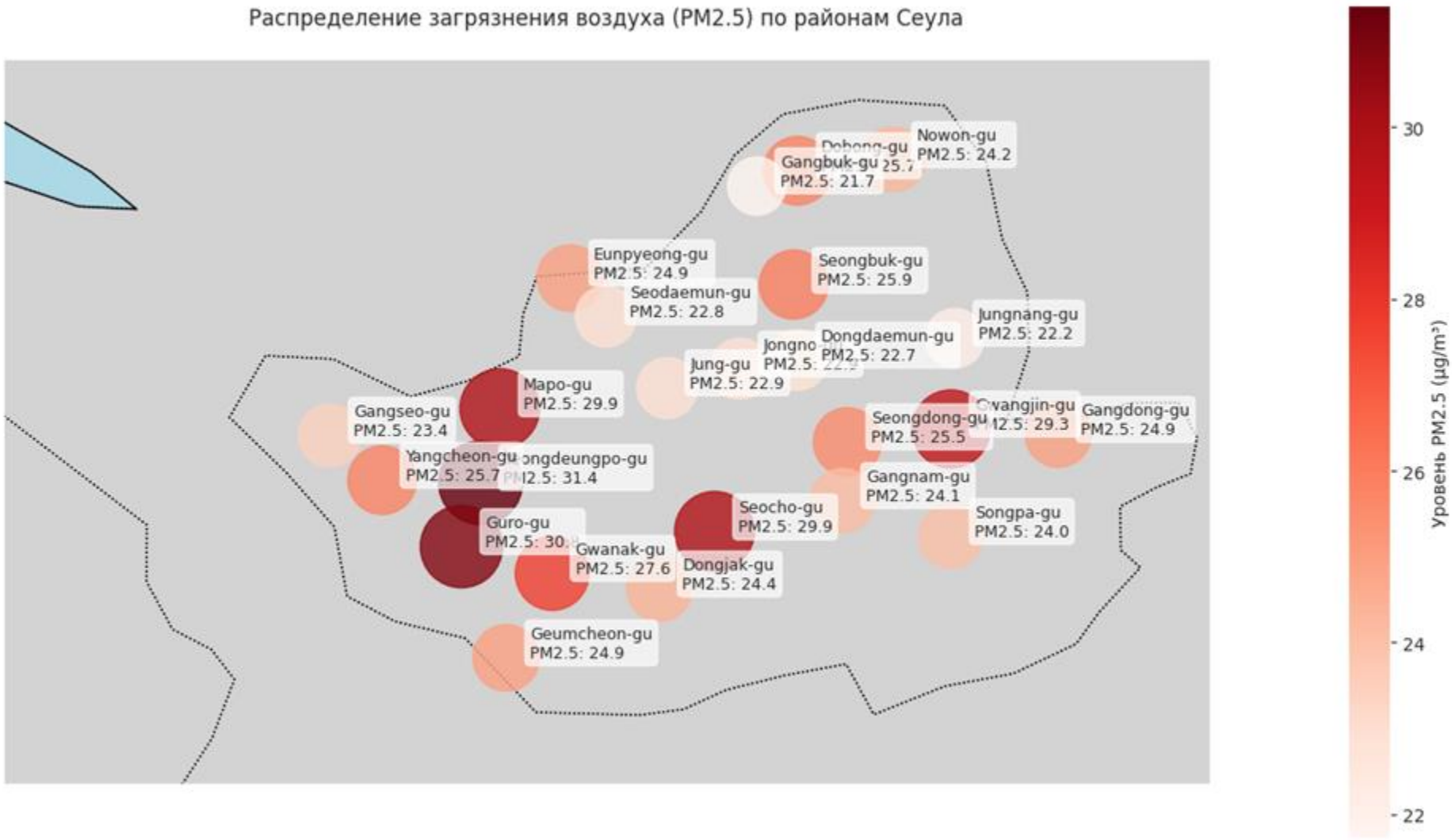

PM2.5
Good(Blue)
Normal(Green)
Bad(Yellow)
Very bad(Red)
Bounded res
PM2.5
0
20
40
60
80
100
120
140
160
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
2019-05
2019-09
Date

Географический анализ загрязнения воздуха в Сеуле

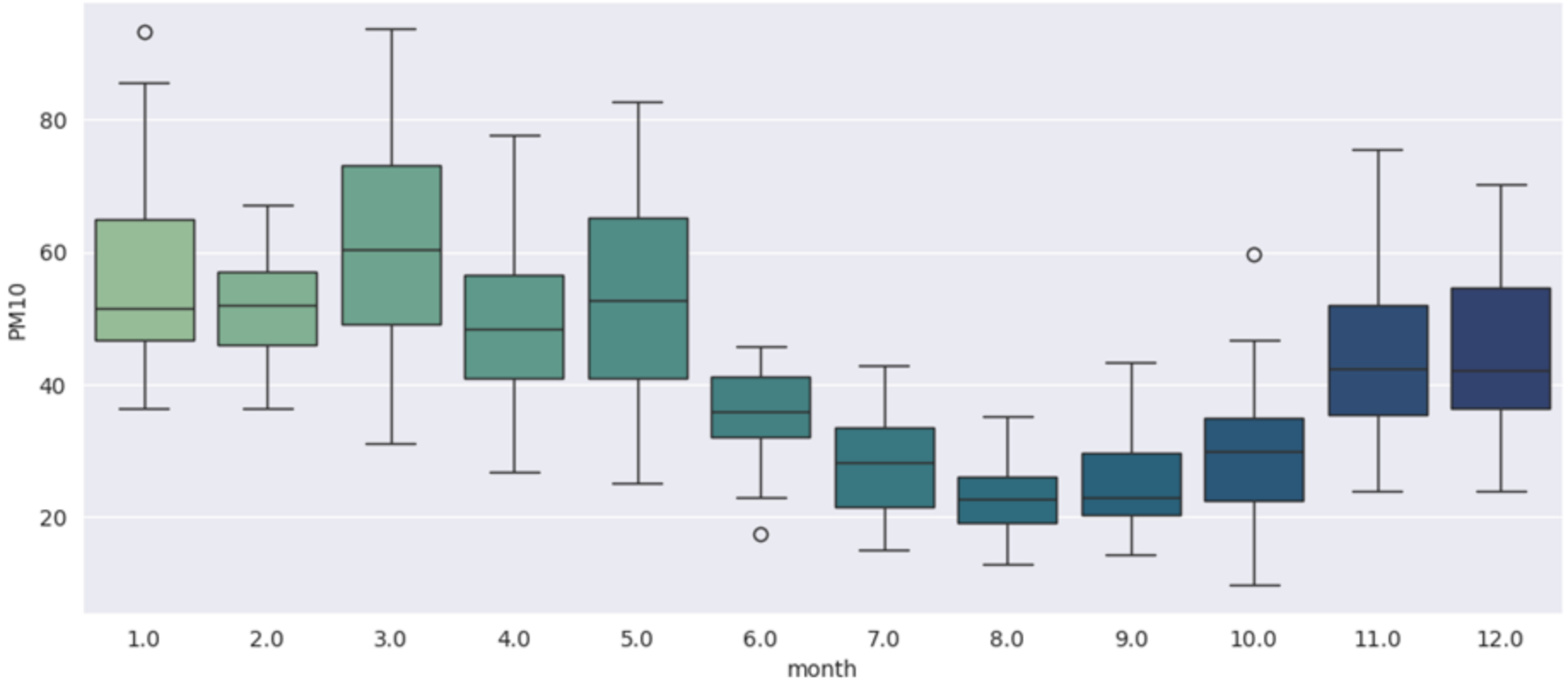


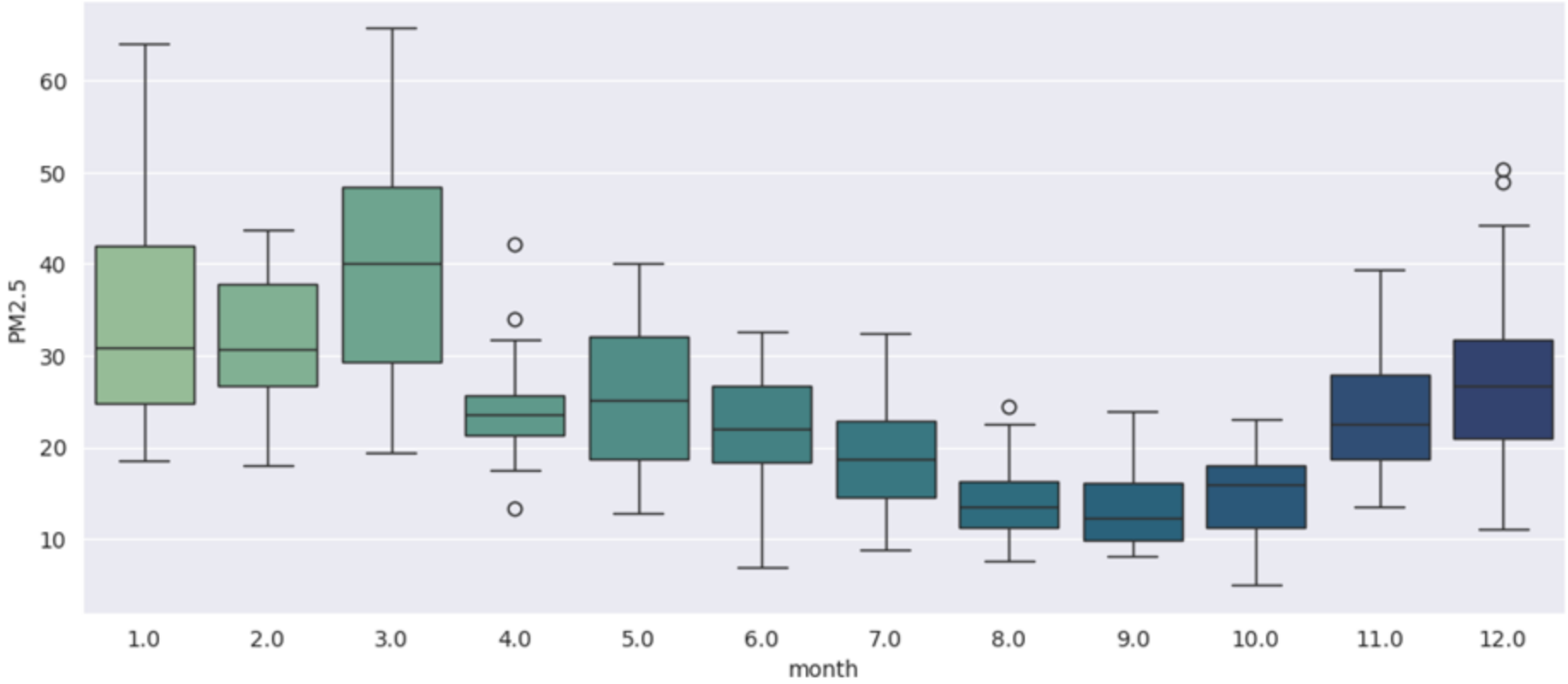

Средний уровень PM10 по районам
Gangseo-gu
Seocho-gu
Guro-gu
Yeongdeungpo-gu
Seongdong-gu
Mapo-gu
Songpa-gu
Seongbuk-gu
Gwangjin-gu
Gangdong-gu
Gwanak-gu
Dobong-gu
Yangcheon-gu
Eunpyeong-gu
Dongjak-gu
Seodaemun-gu
Nowon-gu
Geumcheon-gu
Gangnam-gu
Gangbuk-gu
Dongdaemun-gu
Jung-gu
Jongno-gu
Jungnang-gu
0
10
20
30
40
50
PM10 (μg/m³)

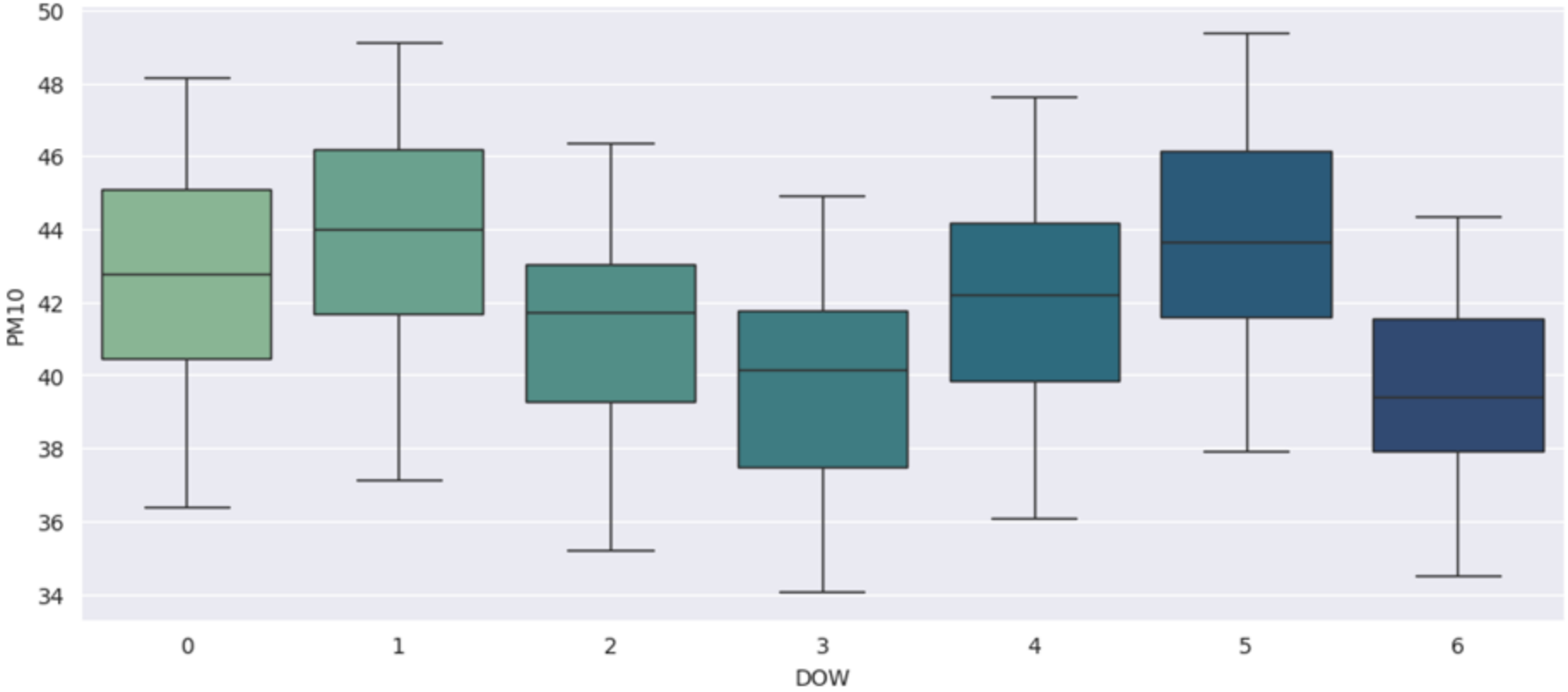

Процент превышений нормативов по районам
Район
Gwanak-gu 23.6 30.3
Yeongdeungpo-gu 23.4 35.2
Mapo-gu 22.4 29.4
Yangcheon-gu 21.5 28.1
Guro-gu 20.6 30.3
Seongdong-gu 20.1 31.8
Gwangjin-gu 19.6 26.9
Seocho-gu 19.4 31.1
Gangdong-gu 19.0 32.6
Gangnam-gu 18.8 25.0
Dongjak-gu 18.7 26.2
Seodaemun-gu 18.5 27.9
Eunpyeong-gu 18.4 27.6
Nowon-gu 18.3 25.2
Dobong-gu 18.3 26.9
Seongbuk-gu 18.2 30.8
Geumcheon-gu 18.2 23.7
Gangseo-gu 17.7 30.5
Songpa-gu 17.2 28.2
Dongdaemun-gu 16.7 24.5
Jungnang-gu 16.4 23.8
Jung-gu 16.4 22.1
Jongno-gu 16.0 21.8
Gangbuk-gu 15.9 23.7
PM25_exceed
PM10_exceed
Показатель
% превышений нормы
35.0
32.5
30.0
27.5
25.0
22.5
20.0
17.5

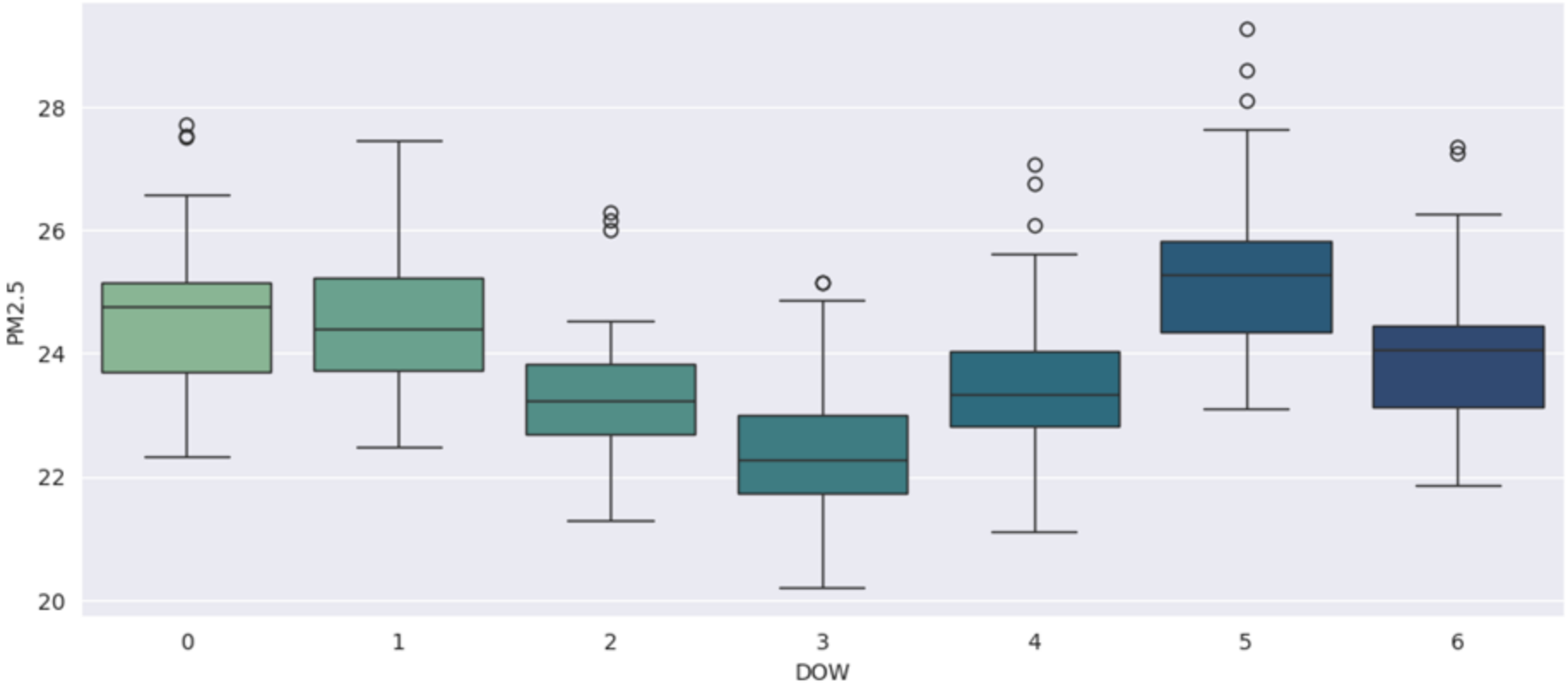

Распределение загрязнения воздуха (PM2.5) по районам Сеула
Nowon-gu
PM2.5: 24.2
Gangbuk-gu
PM2.5: 21.7
Eunpyeong-gu
PM2.5: 24.9
Seongbuk-gu
PM2.5: 25.9
Seodaemun-gu
PM2.5: 22.8
Jungnang-gu
PM2.5: 22.2
Dongdaemun-gu
PM2.5: 22.7
Jung-gu
PM2.5: 22.9
Mapo-gu
PM2.5: 29.9
Gangseo-gu
PM2.5: 23.4
Seongdong-gu
PM2.5: 25.5
Gangdong-gu
PM2.5: 24.9
Yangcheon-gu
PM2.5: 25.7
Gangnam-gu
PM2.5: 24.1
Songpa-gu
PM2.5: 24.0
Seocho-gu
PM2.5: 29.9
Gwanak-gu
PM2.5: 27.6
Dongjak-gu
PM2.5: 24.4
Geumcheon-gu
PM2.5: 24.9
Уровень PM2.5 (μg/m³)
30
28
26
24
22

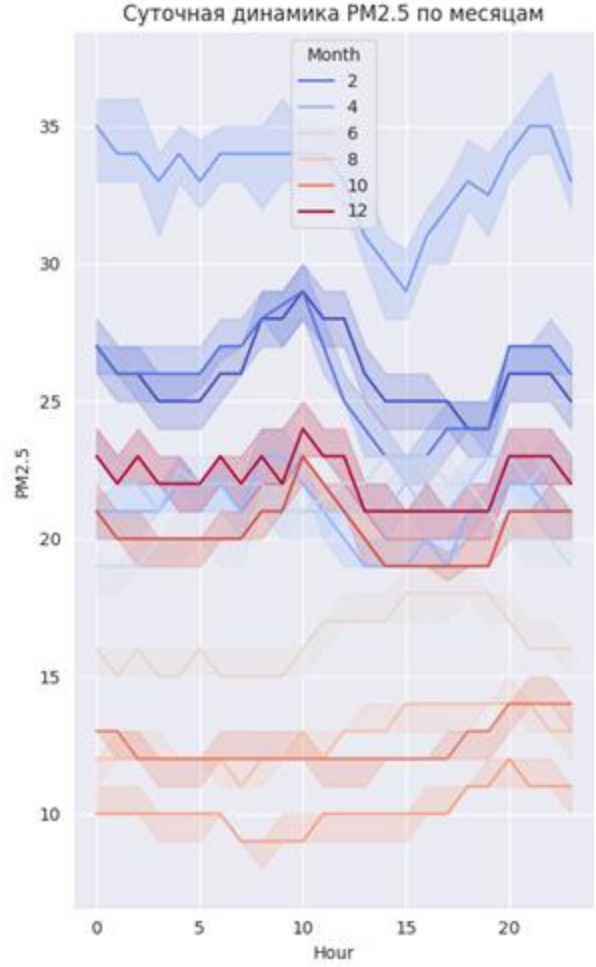
Сезонный анализ загрязнения воздуха в Сеуле

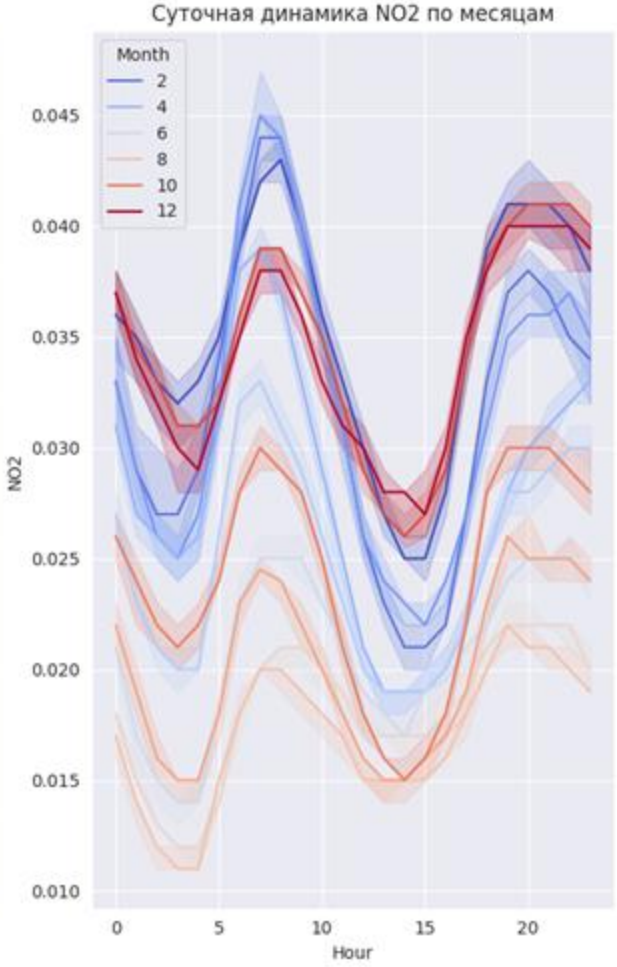


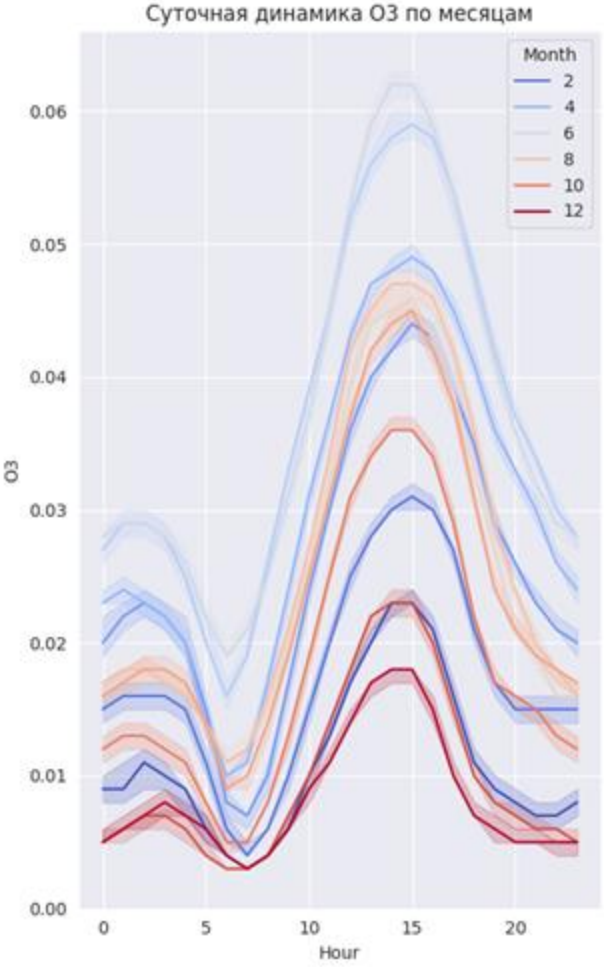

PM2.5
60
50
40
30
20
10
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
month


50
48
46
44
42
40
38
36
34
PM10
0
1
2
3
4
5
6
DOW

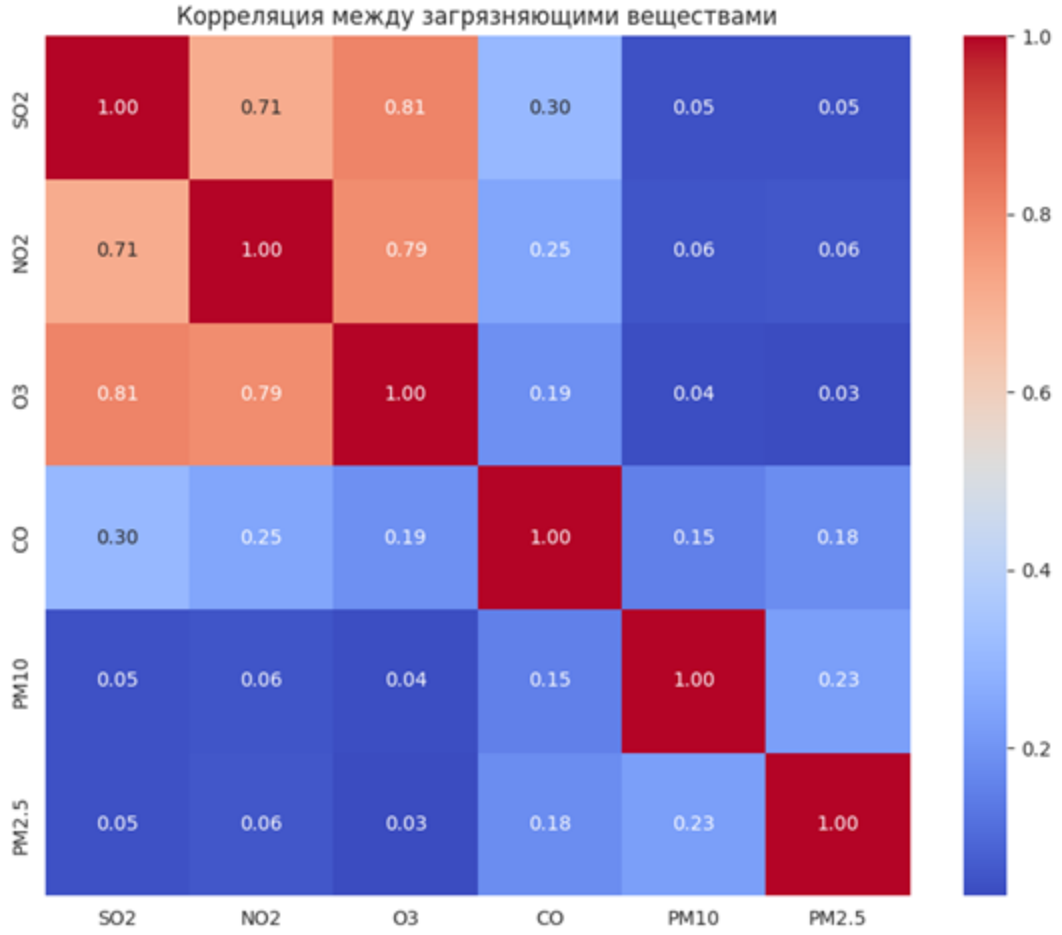

PM2.5
28
26
24
22
20
0
1
2
3
4
5
6
DOW


Суточная динамика PM2.5 по месяцам
Month
2
4
6
8
10
12
PM2.5
Hour


Суточная динамика NO2 по месяцам
Month
2
4
6
8
10
12
NO2
Hour


Суточная динамика O3 по месяцам
Month
2
4
6
8
10
12
O3
Hour

Корреляций между загрязняющими веществами



Задача регрессии

Постановка задачи: предсказание уровня концентрации мелкодисперсных частиц P.M.2.5

Используемые метрики:

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

MSE:

Преимущества:

- Чувствительна к большим ошибкам
- Дифференцируема

Недостатки:

- Чрезмерно штрафует выбросы

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MAE:

Преимущества:

- Устойчива к выбросам
- Результат в исходных единицах измерения

Недостатки:

- Не дифференцируема в нуле
- Менее чувствительна к большим ошибкам, чем MSE

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R2:

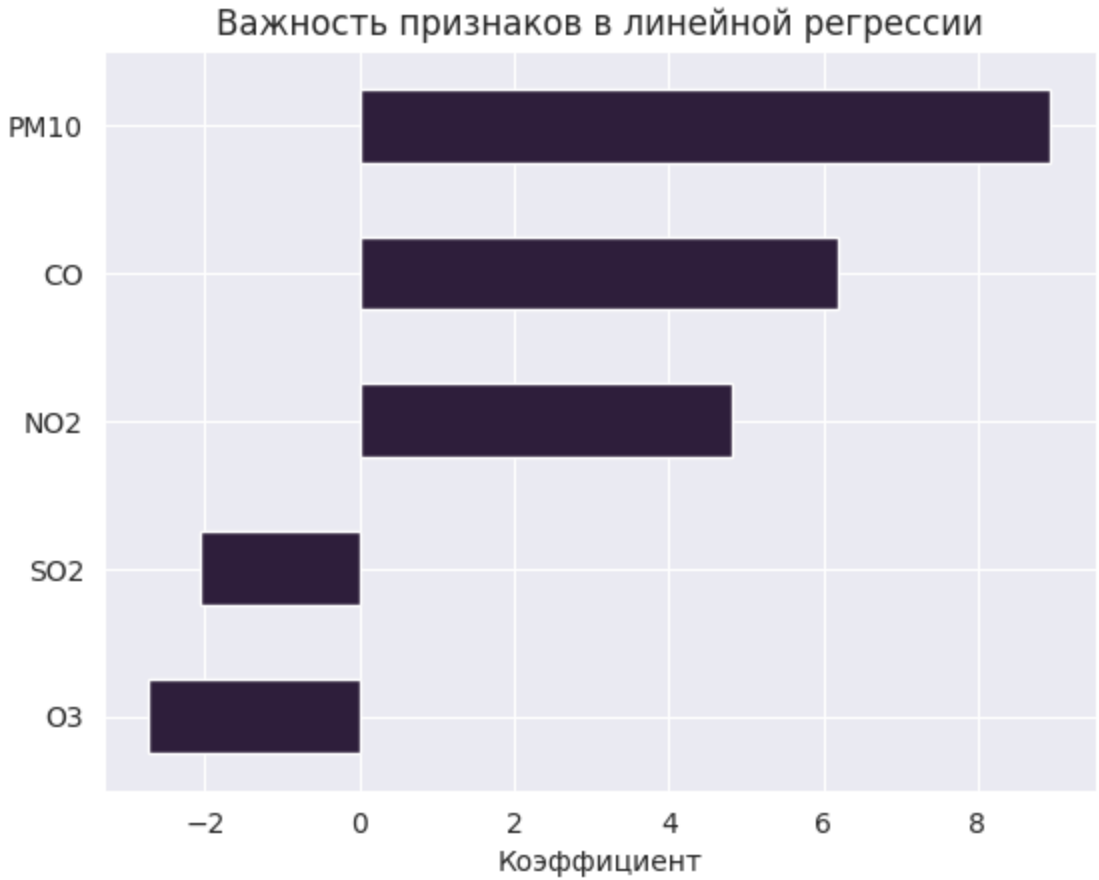
Преимущества:

- Интерпретируема как процент объясненной вариации

Недостатки:

- Может быть отрицательной для плохих моделей
- Чувствительна к выбросам

Линейная регрессия



Линейная регрессия

Работает быстро, но уступает по точности для сложных данных

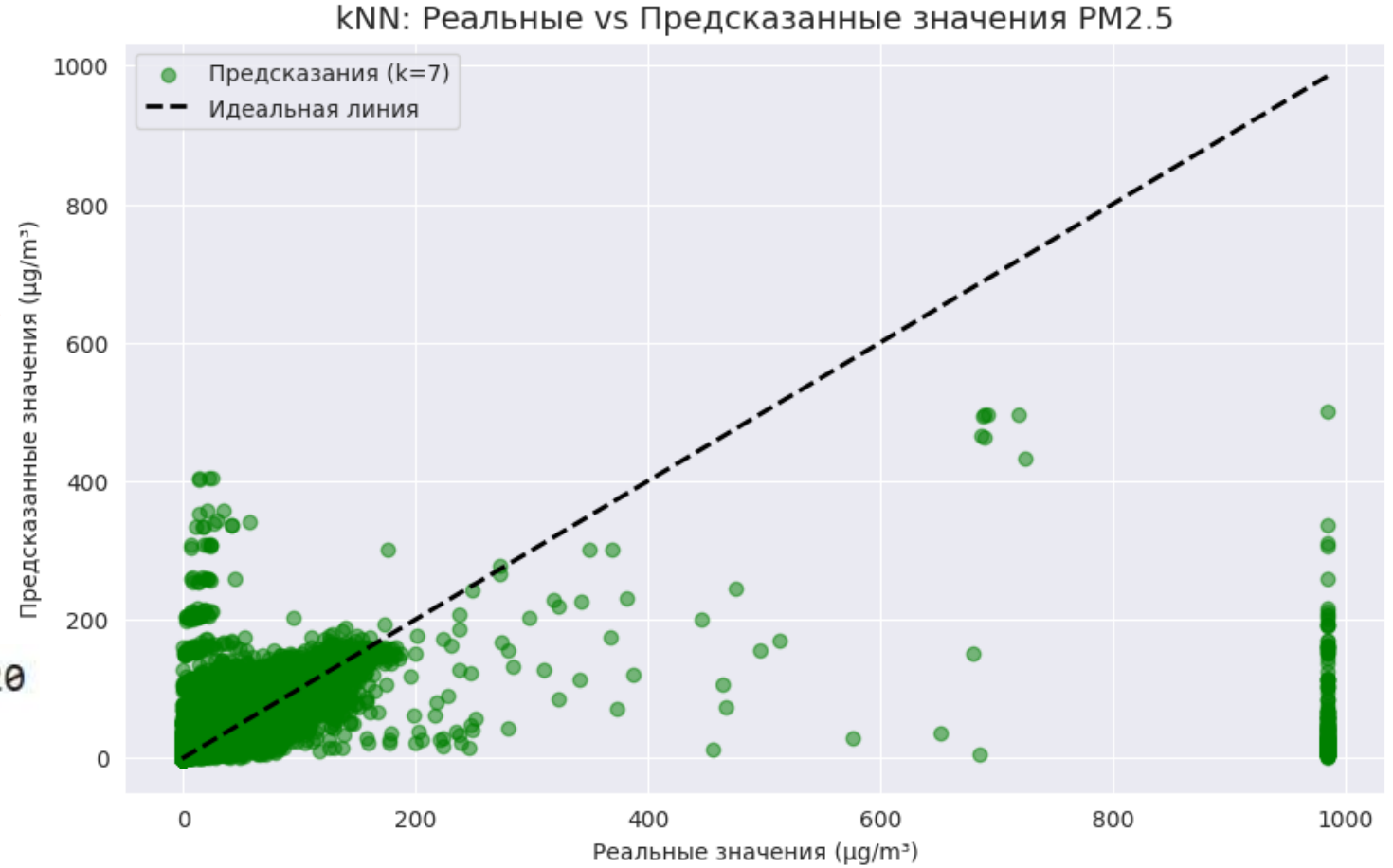
Хорошо интерпретируемые коэффициенты

MSE: 1749.66
MAE: 12.10
R^2: 0.07

kNN

Показывает улучшение качества предсказаний, но демонстрирует характерные "ступеньки" в прогнозах из-за природы метода.

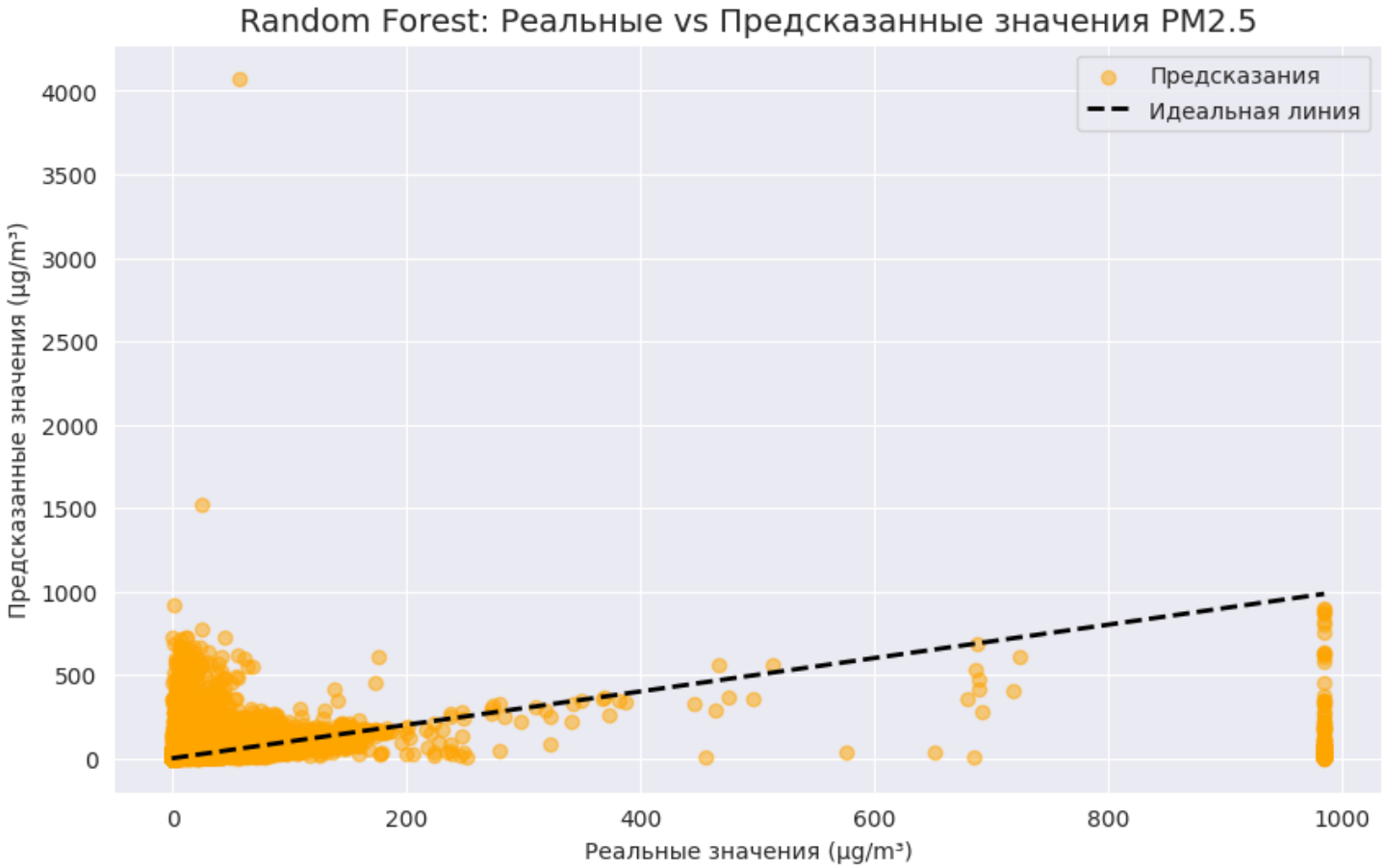
```
Оптимальное k: 20
MSE: 1611.41
MAE: 8.55
R²: 0.14
```



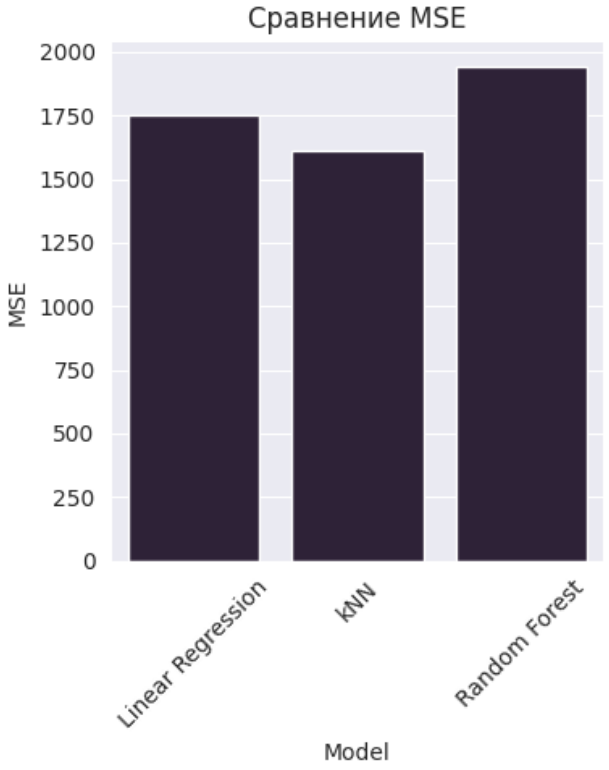
Random Forest

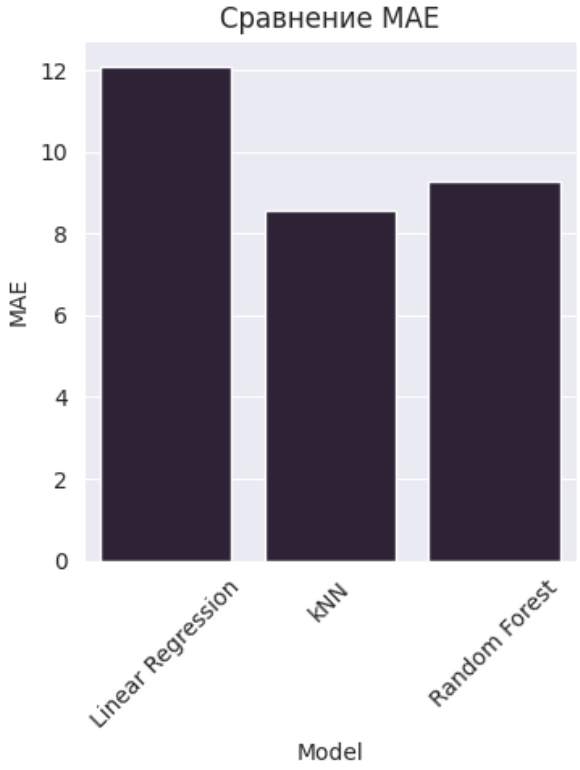
Обеспечивает наиболее точные предсказания во всем диапазоне значений, минимальное систематическое смещение и устойчивость к выбросам

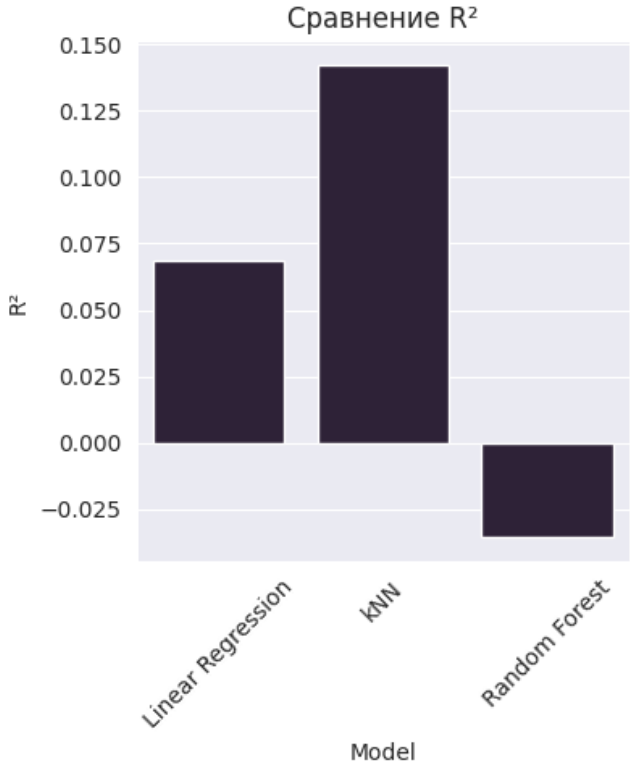
```
Random Forest:
MSE: 1944.89
MAE: 9.28
R²: -0.04
```

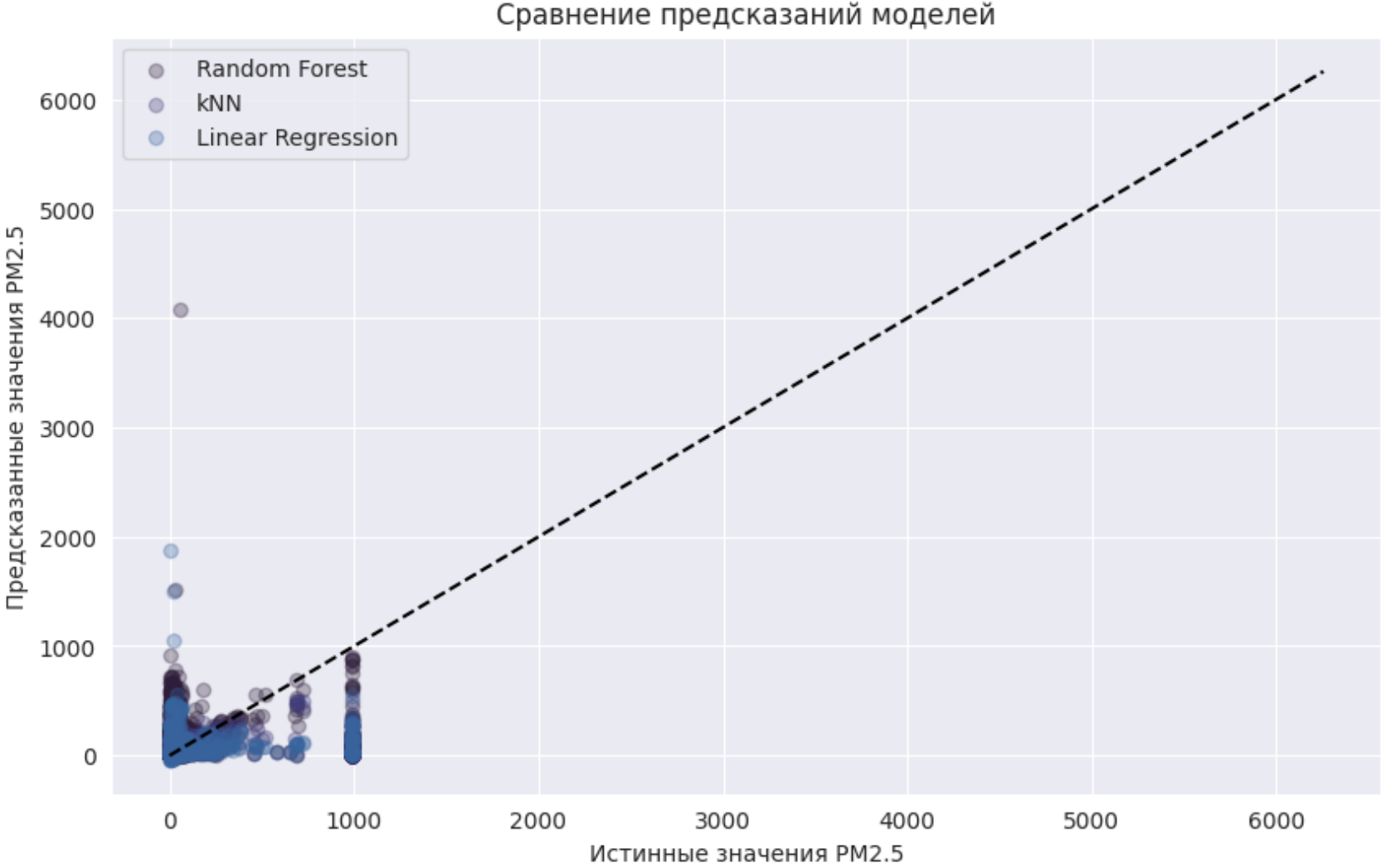


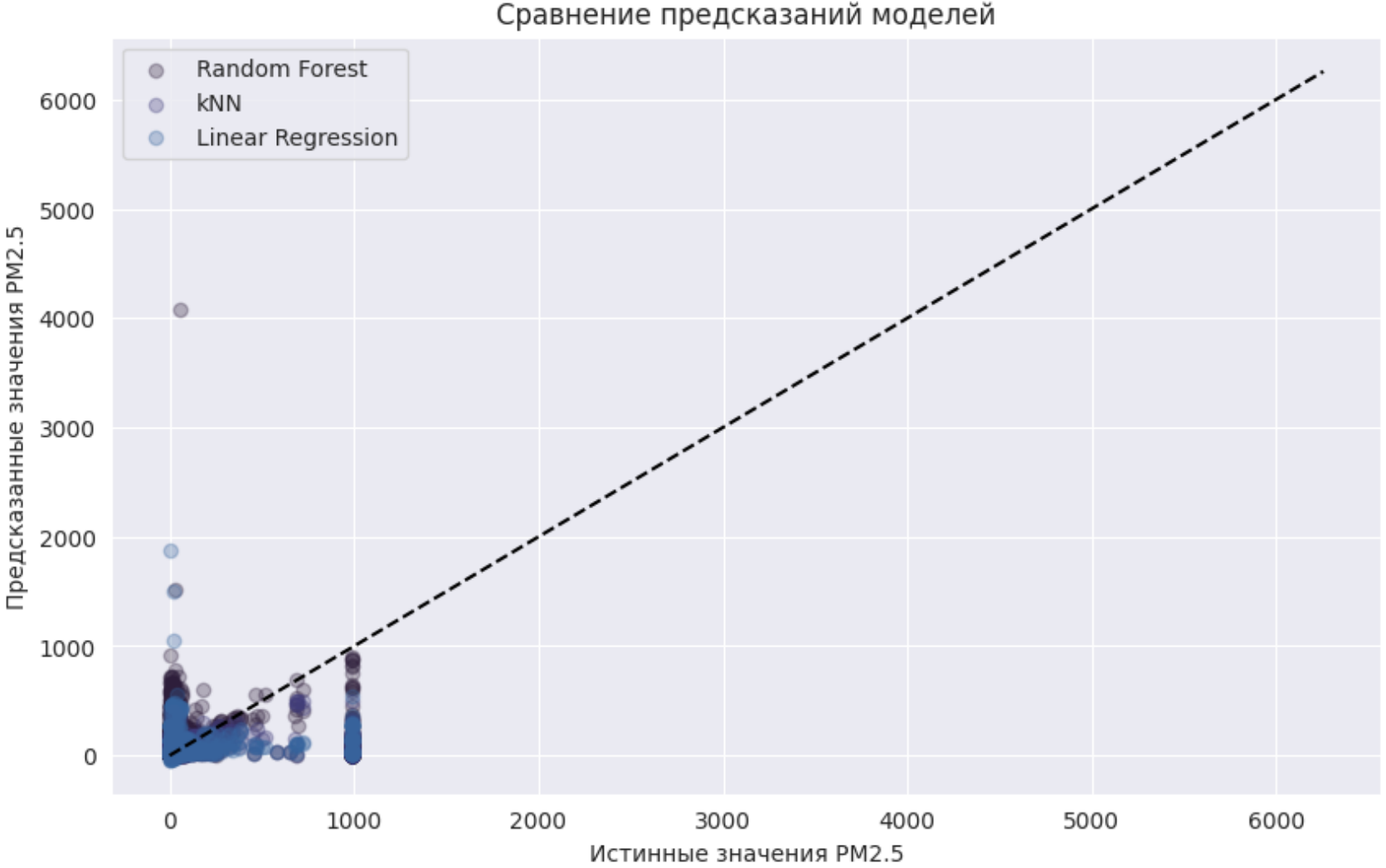
Результаты сравнения








Сравнение предсказаний моделей
Random Forest
kNN
Linear Regression
Предсказанные значения PM2.5
Истинные значения PM2.5
0
1000
2000
3000
4000
5000
6000

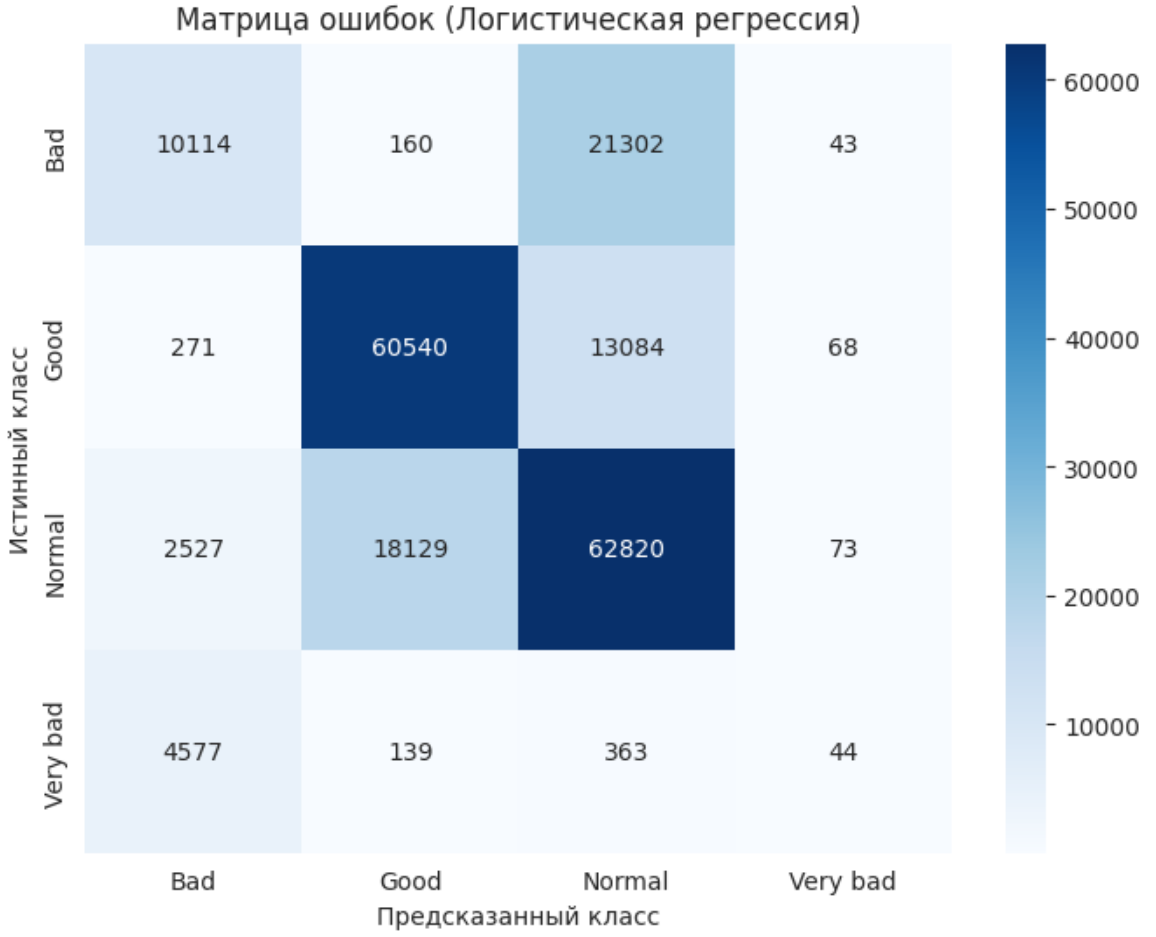

Сравнение предсказаний моделей
Random Forest
kNN
Linear Regression
Предсказанные значения PM2.5
Истинные значения PM2.5
0
1000
2000
3000
4000
5000
6000

ЗАДАЧА КЛАССИФИКАЦИИ

Автоматически определять уровень опасности загрязнения воздуха частицами PM2.5, относя каждое наблюдение к одному из заранее определённых классов качества воздуха.

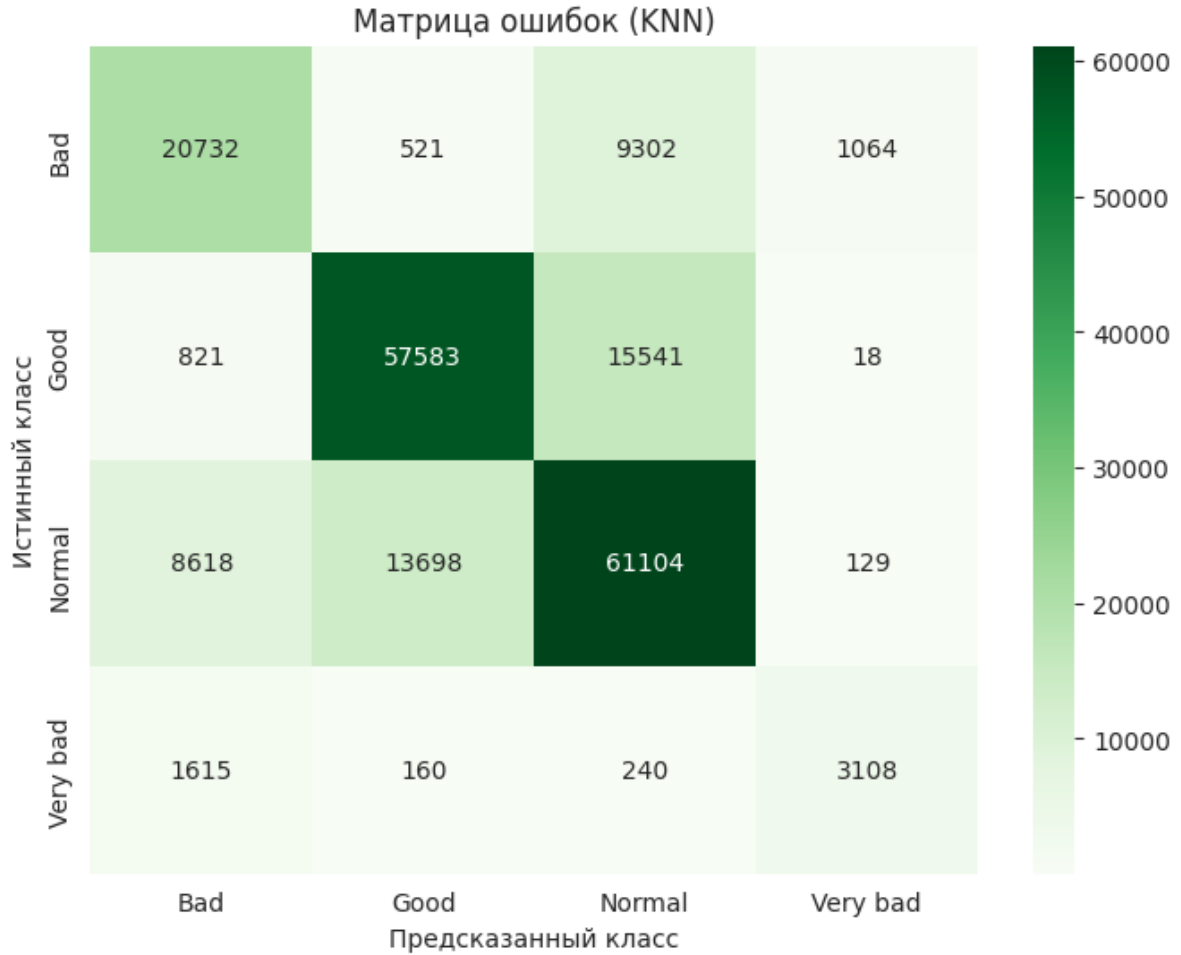
Логистическая регрессия

```
=== Логистическая регрессия ===
Accuracy: 0.6873371976896229
Precision (macro): 0.545445128703609
Recall (macro): 0.4747177865187723
F1-Score (macro): 0.47844374500992226
ROC-AUC (ovr): 0.8806502193156305
```



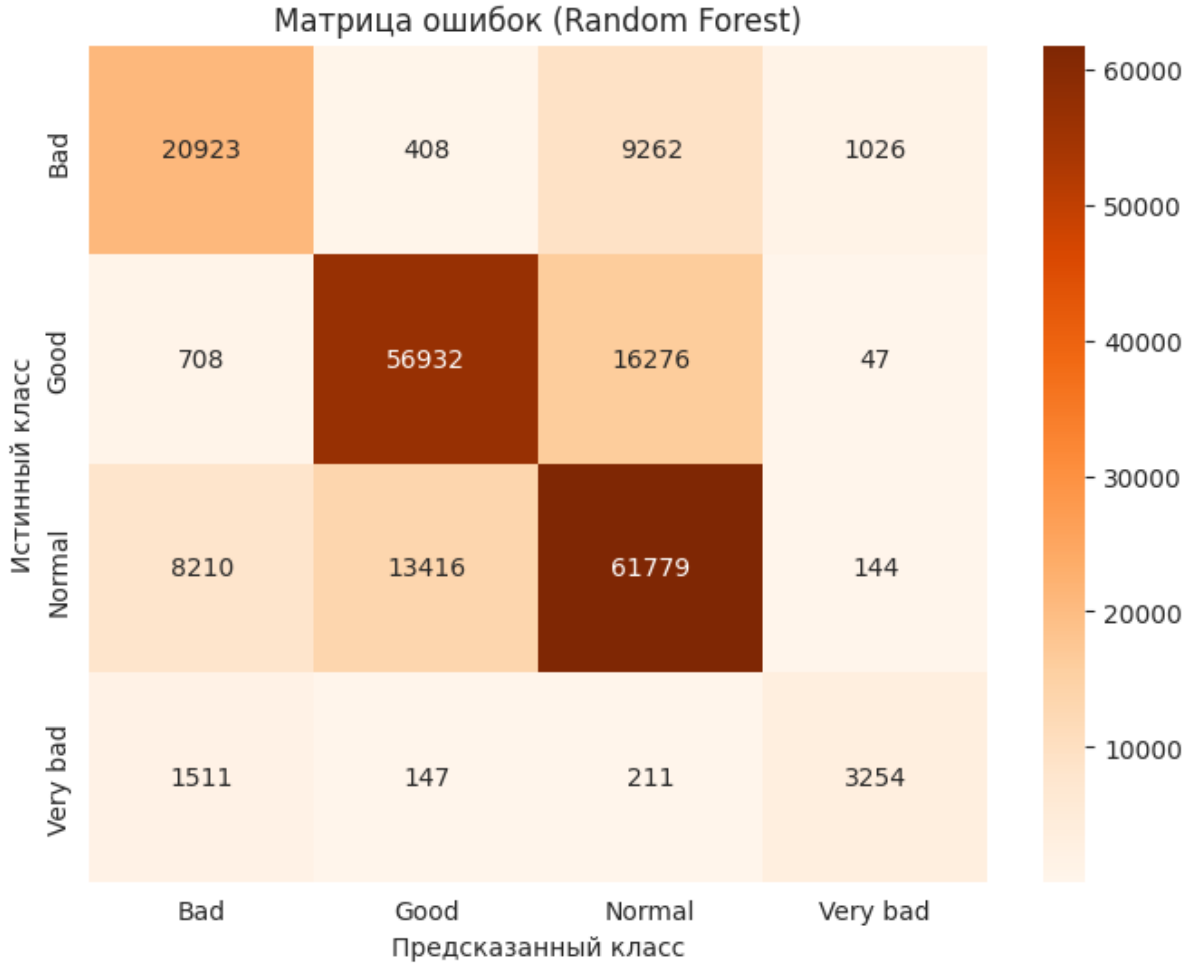
kNN

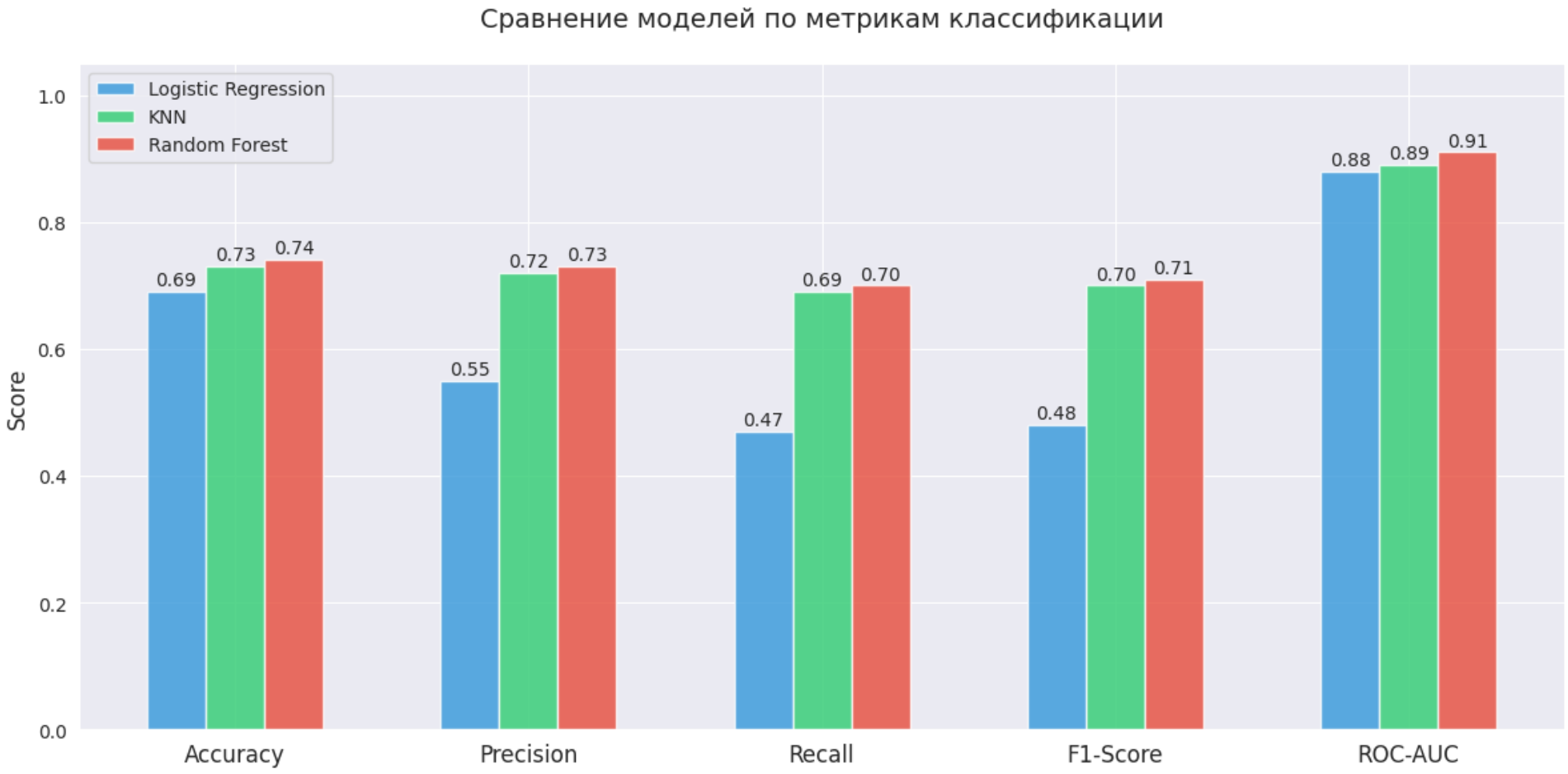
```
=== KNN (k=5) ===
Accuracy: 0.7337146210631441
Precision (macro): 0.7202510236417956
Recall (macro): 0.6930626641225701
F1-Score (macro): 0.7053730398463454
ROC-AUC (ovr): 0.8860601153780329
```



Random Forest

```
=== Random Forest ===
Accuracy: 0.7355730126535361
Precision (macro): 0.7259837835829319
Recall (macro): 0.7015169119534753
F1-Score (macro): 0.7127750663351549
ROC-AUC (ovr): 0.9121539789863153
```




Сравнение моделей по метрикам классификации
Logistic Regression
KNN
Random Forest
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.69
0.73
0.74
0.55
0.72
0.73
0.47
0.69
0.70
0.48
0.70
0.71
0.88
0.89
0.91
Accuracy
Precision
Recall
F1-Score
ROC-AUC

Спасибо за внимание!