AQI

Air Quality index

Куратор – Кирилл Козлов

Студент – Анастасия Довгаль

Признаки

- AQI — Значение индекса качества воздуха (min:6, max:500)
- CO — Значение угарного газа (min:0, max:133)
- Ozone — Значение озона (min:0, max:235)
- NO_2 — Значение диоксида азота (min:0, max:91)
- PM2.5 — Значение твердых частиц (min:0, max:500)

Подготовка данных

Проблема: Отсутствие данных

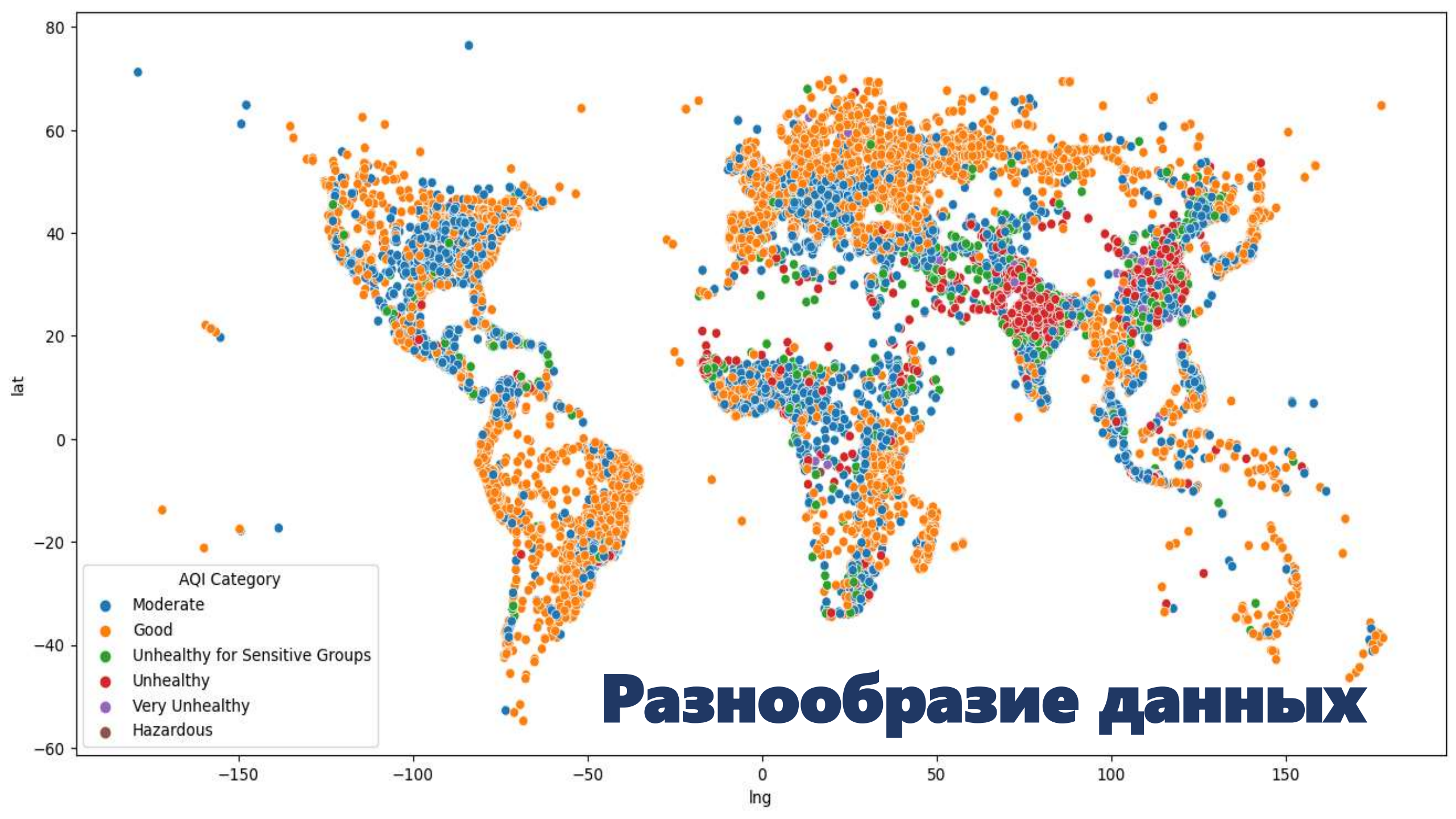
В датасете не хватало данных для анализа его полноты. Датасет был дополнен координатами городов с помощью вспомогательного датасета, а также с помощью библиотеки geopandas.

Проблема: Ошибки в названиях стран

В датасете присутствовали некорректные названия стран, что усложняло визуализацию. Была применена корректировка названий с помощью сервиса geolocator имеющихся координат.

Итог:

Было удалено менее 1% данных после всех обработок. И мы можем визуализировать разнообразие данных.


AQI Category
Moderate
Good
Unhealthy for Sensitive Groups
Unhealthy
Very Unhealthy
Hazardous
lat
lng
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150
Разнообразие данных

EDA

Проблема: Наличие выбросов

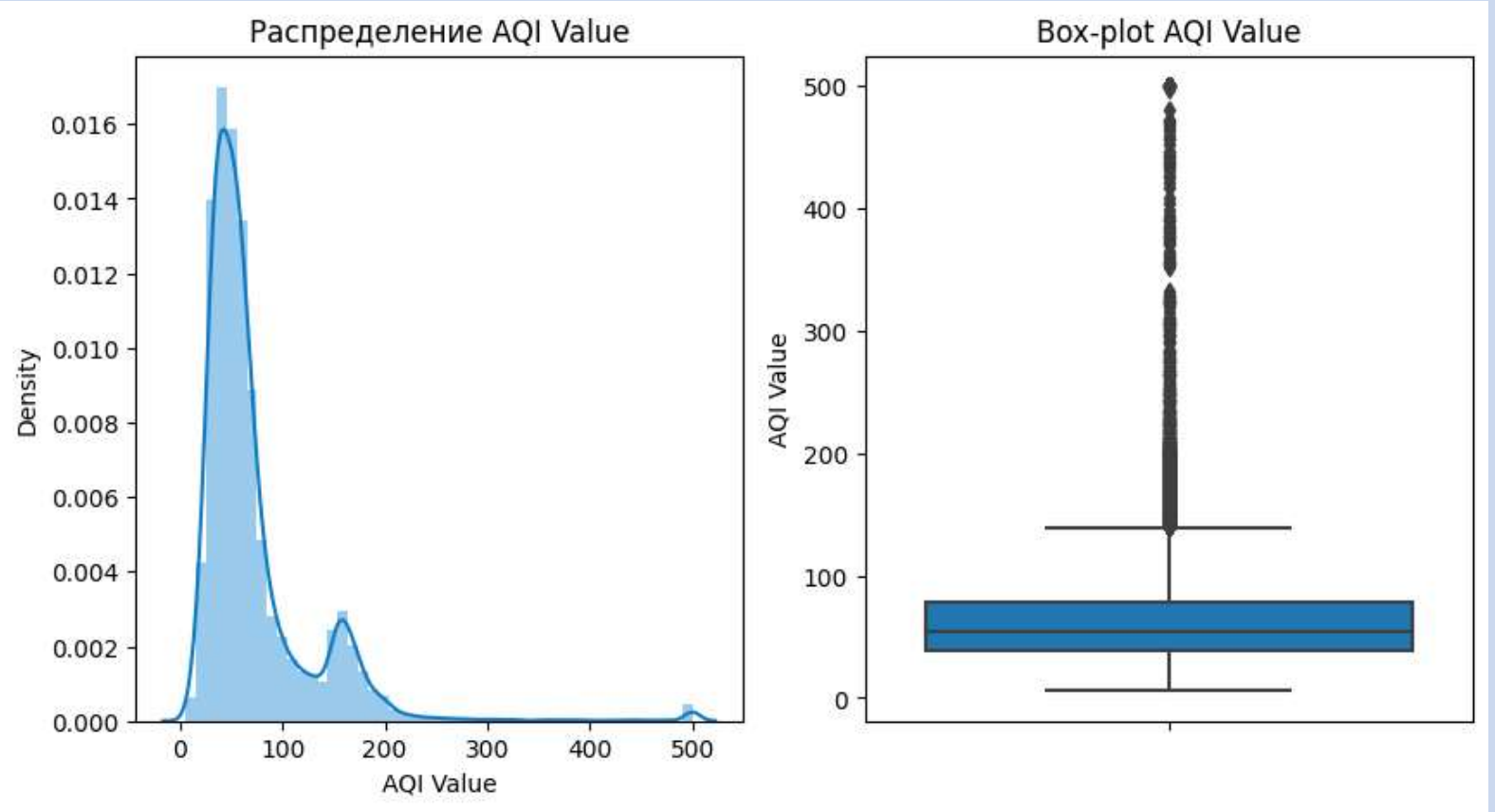
В датасете целевая переменная распределена неравномерно, для оценки объема выбросов использовался фильтр Хэмпеля с 3 стандартными отклонениями. Но таких строк оказалось более 3500 (~13%).

Проблема: Поиск выбросов

После оценки с помощью фильтра Хэмпеля были выбраны строки выходящие за рамки 0.975 квантиля и при этом по всем показателям, кроме PM2.5, стоят относительно низкие значения.

Итог:

Было удалено менее 1% данных после всех обработок. Мы сохранили разнообразие данных и удалили те строки, в которых были данные несоизмеримые с реальностью.



Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category	countryAverage
India	Gohana	500	Hazardous	1	Good	47	Good	1	Good	500	Hazardous	153,83
India	Pilkhuwa	500	Hazardous	2	Good	61	Moderate	2	Good	500	Hazardous	153,83
United States	Durango	500	Hazardous	133	Unhealthy for Sensitive Groups	0	Good	53	Moderate	500	Hazardous	59,60

Корреляция с AQI

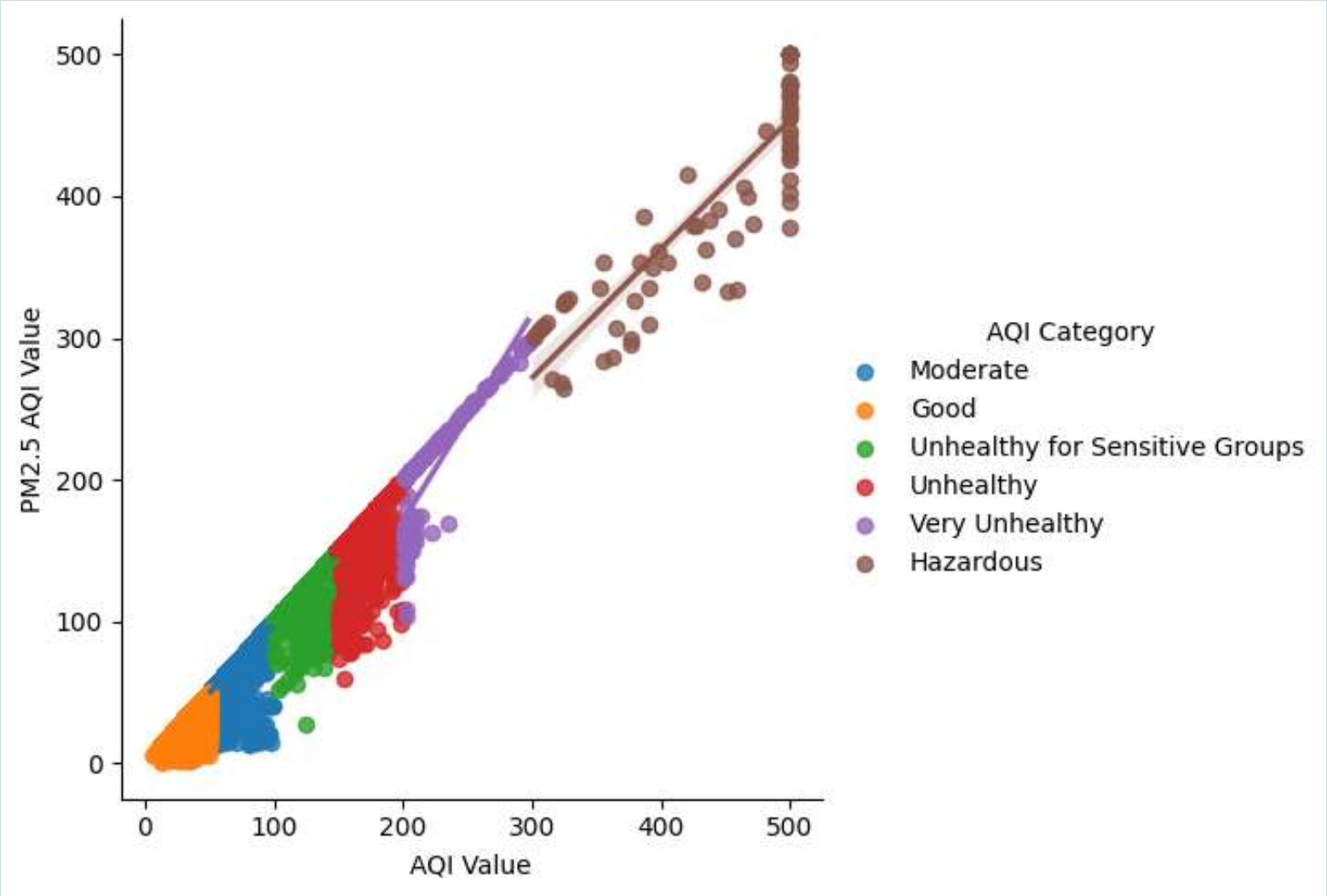
60%
Угарный газ (CO)

44%
Озон (O_3)

28%
Диоксид азота (NO_2)

98%
Твердые частицы ($PM_{2.5}$)

Корреляция между целевой переменной AQI и PM2.5



ML

Задача регрессии

Для решение нашей задачи важно минимизировать возможную ошибку в предсказаниях, поэтому на всех моделях оценивалась MSE.

Простые модели

Так как данные сильно скоррелированы с целевой переменной, то начать было решено с линейной регрессии.

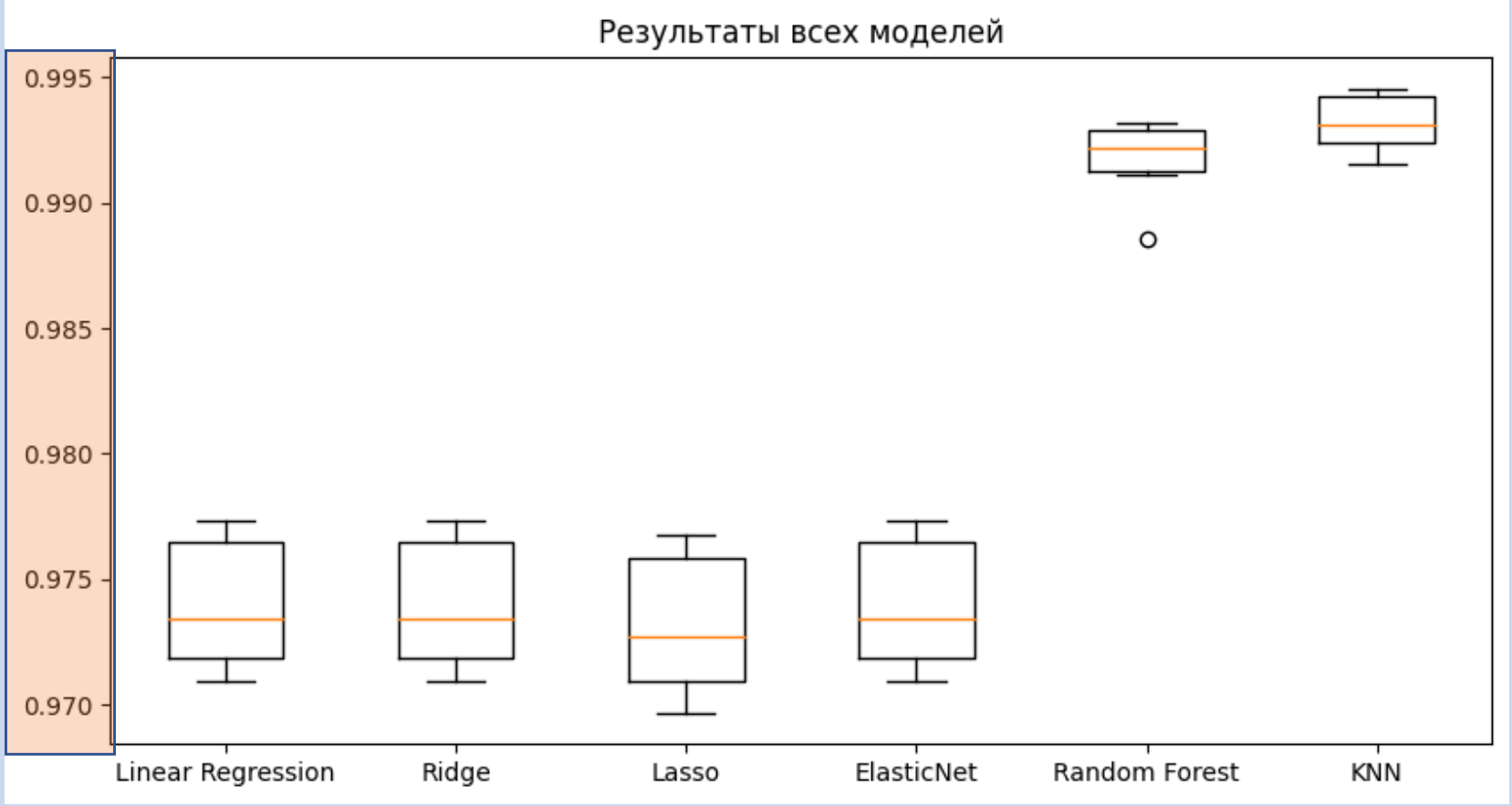
Итог:

Качество на простых моделях было неудовлетворительным. Поэтому были проведены эксперименты с более сложными: RandomForestRegressor, KNeighborsRegressor.

Результаты тестирования

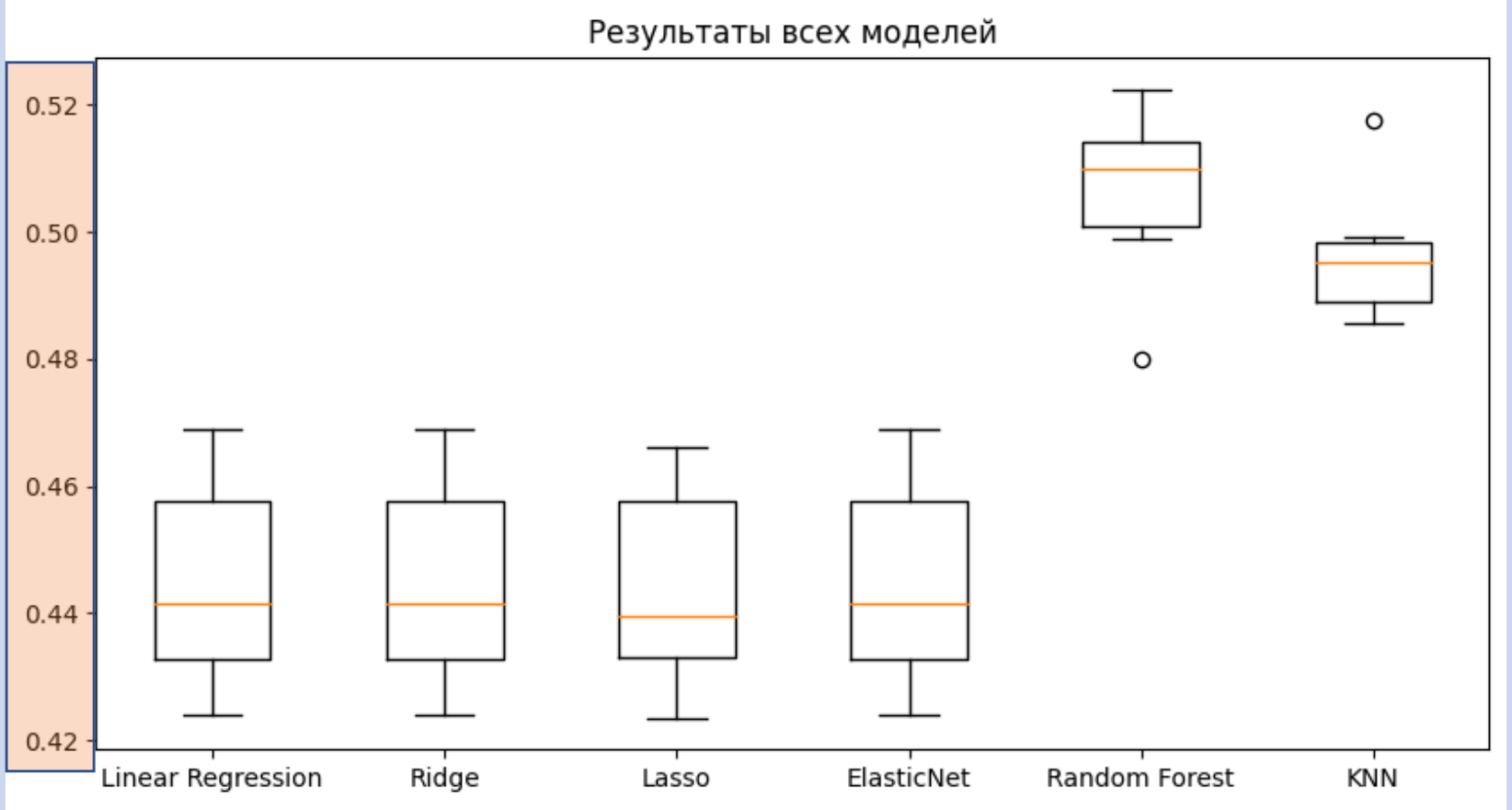
Модель	Train	Test
LinearRegression	58.55	64.71
Lasso	60.33	66.73
ElasticNet	58.55	64.73
RandomForestRegressor	16.12	19.43
KNeighborsRegressor	2.83	4.41

Результаты кросс-валидации на вещественных признаках



- Linear Regression
- Ridge(alpha=0.1)
- Lasso
- Elastic Net(alpha=0.0004, l1_ratio=0.1)
- RandomForest Regressor(n_estimators=100, max_depth=5)
- Kneighbors Regressor(n_neighbors=10)

Результаты кросс-валидации без признака PM2.5



- Linear Regression
- Ridge(alpha=0.1)
- Lasso
- Elastic Net(alpha=0.0004, l1_ratio=0.1)
- Random Forest Regressor(n_estimators=100, max_depth=5)
- Kneighbors Regressor(n_neighbors=10)

Streamlit визуализация

Настройки

Страна

China

Город (Страна - China)

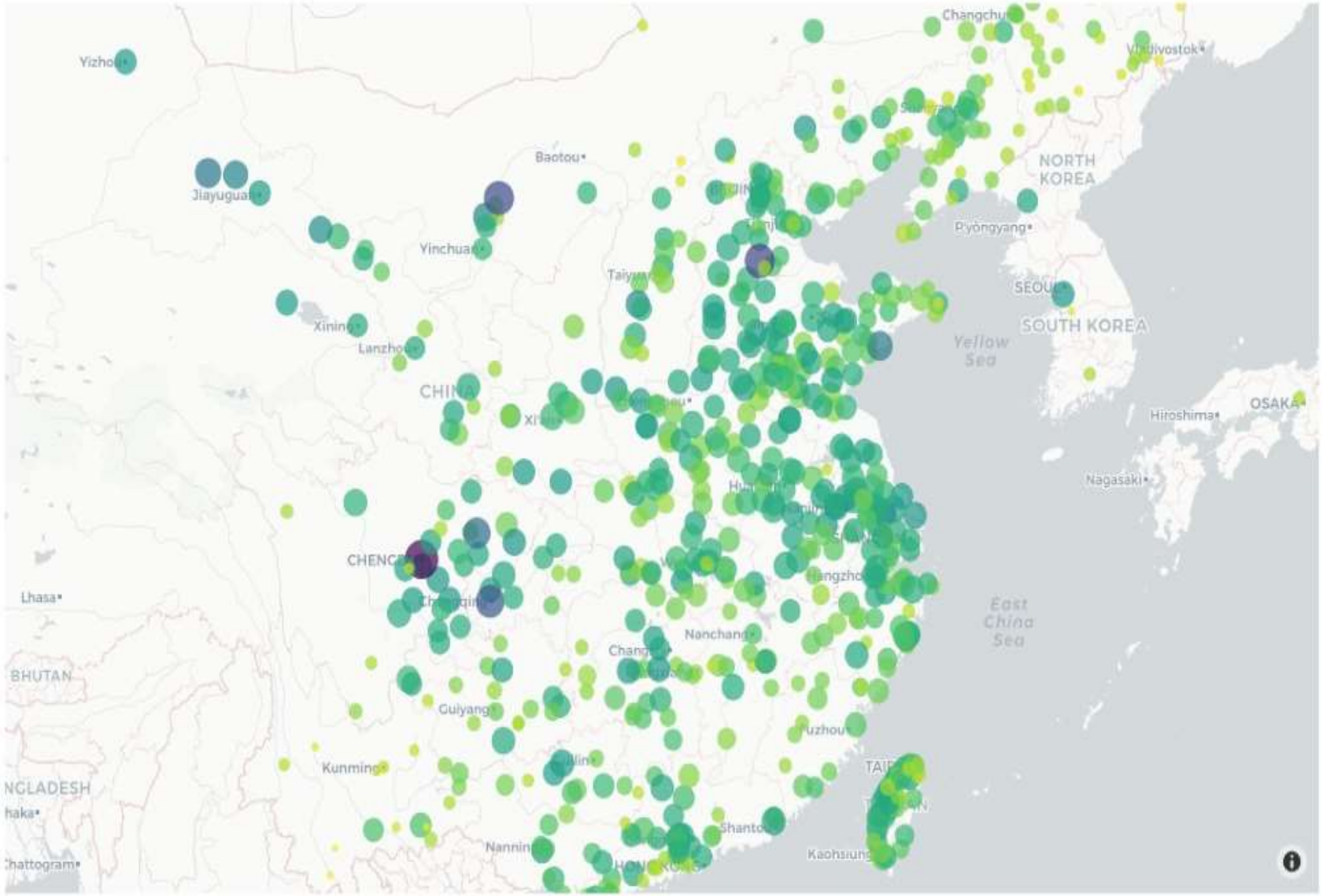
-

Размер меток

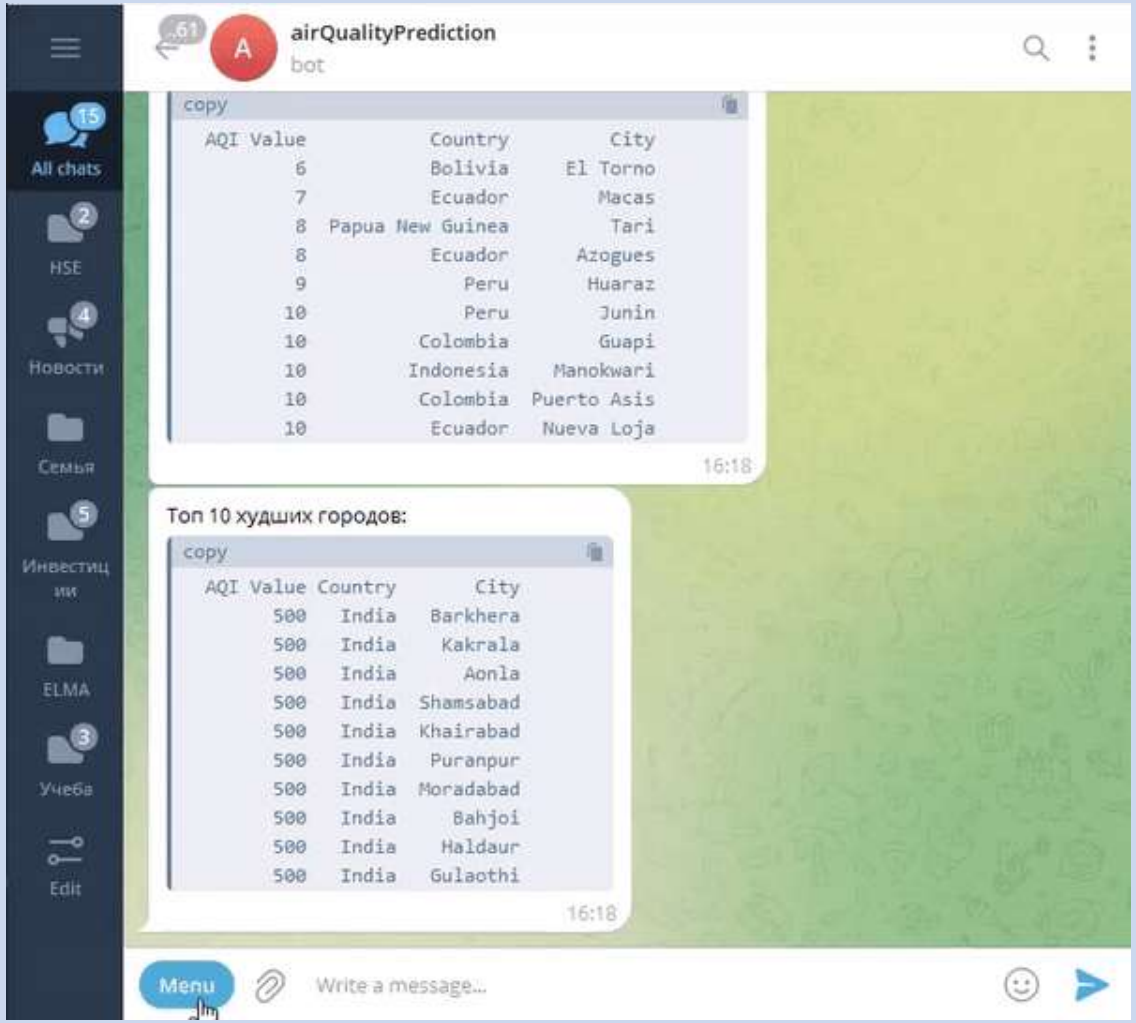
AQI Value

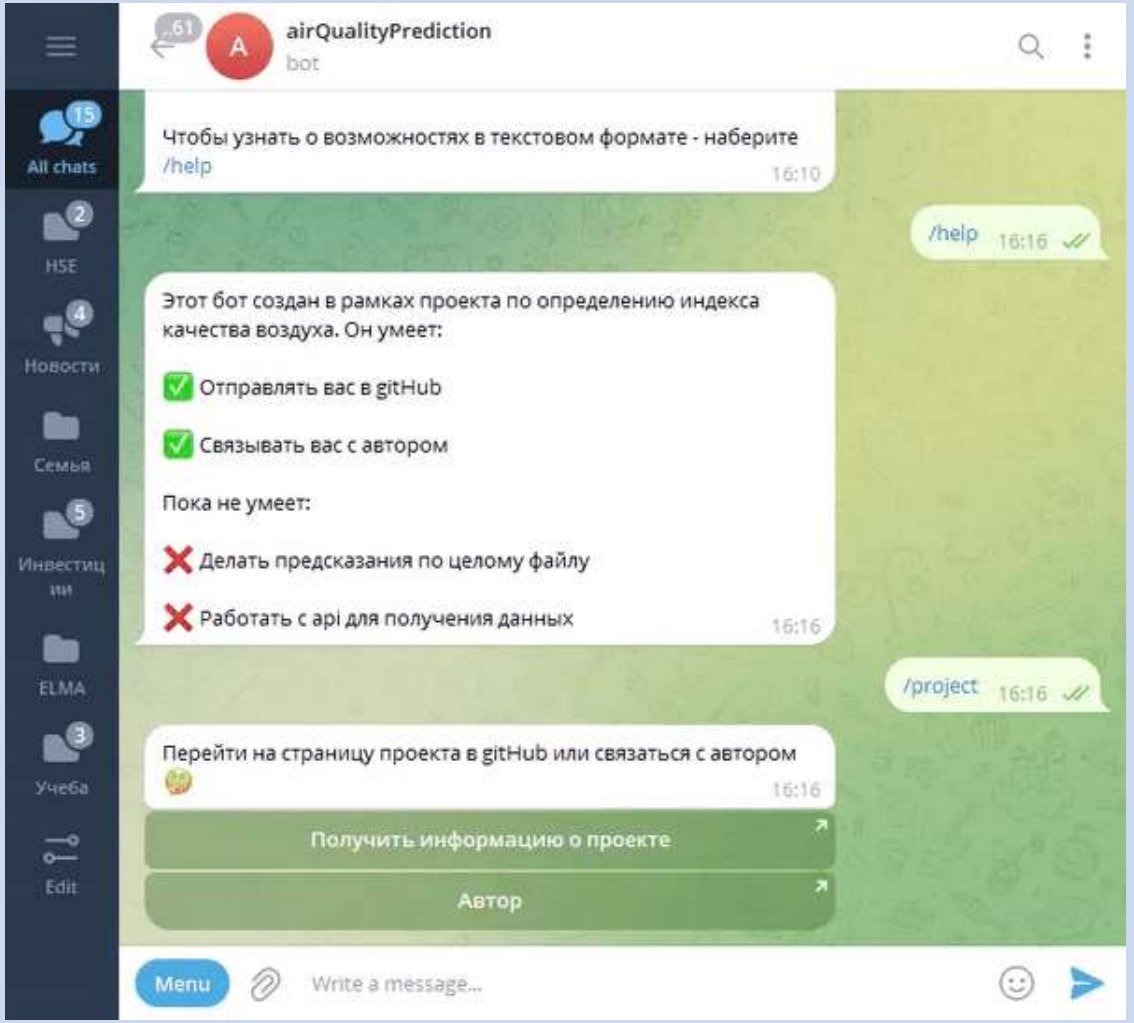
Цвет меток

PM2.5 AQI Value



Пример работы TgBot





Грядущие планы

- Добавить к сервисам **(Streamlit** и **TGBot)** возможность получения данных из https://open-meteo.com/
- Изучить данные с годовыми показателями
- Делать предсказания на период времени
- Обучить нейросеть без признака **PM2.5**