Задание для кандидата на позицию аналитика данных

Контекст задачи
Органы экологического мониторинга в крупных городах собирают большое количество данных о качестве воздуха. Эти данные включают показатели концентрации загрязняющих веществ и метеорологические параметры. Ваша задача — проанализировать данные о качестве воздуха в течение трех лет в нескольких городах, выявить аномальные изменения в показателях загрязнения и предложить возможные объяснения для этих аномалий.

Цель задания
Применить методы анализа данных для обнаружения аномалий и отклонений в показателях качества воздуха, используя соответствующие алгоритмы. Ваша задача — найти выбросы, выявить потенциальные причины аномалий и предложить возможные шаги для улучшения экологического мониторинга.

Данные
Вам необходимо выбрать и использовать один из наборов данных, предоставленных ниже:

1. OpenAQ Platform https://openaq.org/
Данные о качестве воздуха из более чем 100 стран.

2. UCI Machine Learning Repository – Air Quality Data Set
https://archive.ics.uci.edu/ml/datasets/Air+Quality
Данные из Италии за 2004–2005 годы с показателями загрязнения и метеорологическими данными.

Шаги выполнения задания

1. Выберите набор данных из предложенного списка и загрузите его. Обоснуйте свой выбор.

2. Предварительная обработка данных
- Очистите данные от пропусков и аномальных значений, которые могут быть результатом ошибок измерения.
- Проведите исследовательский анализ данных (EDA) и визуализируйте основные тенденции в данных.

3. Поиск аномалий
Используйте один или несколько методов для обнаружения аномалий в данных. Используя, но не ограничиваясь списком:
- Локальная оценка выбросов (Local Outlier Factor, LOF)
- Изолирующий лес (Isolation Forest)
- Метод соседей (k-nearest neighbors) или другие подходы по вашему выбору.

Опишите, почему вы выбрали тот или иной метод, и какие аномалии удалось обнаружить.

4. Интерпретация аномалий
Проанализируйте найденные аномалии. Ответьте на следующие вопросы:
- Какие из них кажутся наиболее значительными?
- В какие периоды или регионы они происходят?
- Могут ли эти выбросы быть вызваны природными или антропогенными факторами?

5. Визуализация результатов
Визуализируйте обнаруженные аномалии с помощью графиков (линейные графики, тепловые карты, диаграммы рассеяния и т.д.). Визуализация должна быть наглядной и легко интерпретируемой.

6. Отчет
Подготовьте отчет (PDF или презентацию), в котором:
- Опишите выбранные данные и методы обработки.
- Представьте и визуализируйте результаты поиска аномалий.
- Предложите возможные объяснения найденных аномалий и их последствия.
- Предложите возможные меры по улучшению мониторинга качества воздуха.

Критерии оценки
- Качество предобработки данных и проведенного EDA.
- Глубина анализа и правильность применения методов обнаружения аномалий.
- Оригинальность и точность интерпретации аномальных событий.
- Четкость и визуальная привлекательность отчетности.
- Способность формулировать обоснованные гипотезы и предложения по улучшению мониторинга.

Документация по методам аномалий
- Local Outlier Factor: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html
- Isolation Forest: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html
- Примеры использования: https://scikit-learn.org/stable/auto_examples/ensemble/plot_isolation_forest.html