

Challenge 3: Help Residents Analyse Data More Easily

Challenge Description

Hollandse Luchten collects large amounts of sensor data on air quality. However, analysing and interpreting this data can be difficult. It often requires specific expertise in air quality, particulate matter, and data analysis.

This challenge seeks solutions that enable residents to explore and analyse air-quality data themselves, even if they do not have this specialist knowledge. The key objective is to make data analysis more accessible while ensuring that results remain valid, understandable, and transparent.

The core idea: Residents ask their questions naturally, the agent/tool interprets the request, uses available data analysis tools to answer it, and provides results along with a complete audit trail showing which steps were taken.

Problem Statement

Residents who want to understand air quality in their area can use existing tools (<https://analyseren.samenmeten.nl/>) to select sensor data, filter possible outliers, and create visualizations such as pollution roses, bar charts, and time-series plots. However, using these tools still requires users to understand data-analysis concepts and make methodological choices.

This challenge is to develop a solution that makes air-quality data analysis more accessible to residents. The solution should help users formulate and investigate their own questions, guide them through appropriate analytical steps, and explain the results in clear language.

The solution should also make the analysis transparent. Users need to be able to see which data was used, which choices were made, which methods were applied, and how these choices may affect the outcome. This enables residents to carry out analyses independently in a valid, understandable, and reproducible way.

A solution may, for example:

1. **Help** residents turn a question into an appropriate analysis.
2. **Support** the selection of relevant sensor data, locations, and time periods.
3. **Guide** users in making methodological choices, such as filtering outliers or comparing groups.
4. Create clear and useful **visualizations**.
5. **Explain** results, assumptions, limitations, and uncertainty in accessible language.
6. Provide a **transparent record** of the data, methods, decisions, and steps used.

7. **Bonus:** Support follow-up questions and refinements.

Possible techniques

Teams may combine one or more of the following techniques:

Large Language Models: Use models such as GPT, Claude, or similar where they add clear value in understanding residents' questions and translating them into an analysis request. If other means than LLMs provide a sufficient solution, this is encouraged. Prefer simpler, less resource-intensive approaches when these meet the requirements with comparable reliability and transparency.

Tool integration: Enable the agent to select and call suitable data-analysis functions, for example for filtering data, comparing groups, or creating visualisations.

Planning and reasoning: Break complex questions into clear analysis steps that can be reviewed and reproduced.

Retrieval-Augmented Generation (RAG): Retrieve relevant information about available datasets, sensor metadata, analysis methods, and tool documentation.

Structured and auditable output: Produce consistent results that show the answer, data used, analysis steps, assumptions, and limitations.

A simple, well-validated approach may be more valuable than a complex solution that is difficult to explain or verify. Teams should be able to substantiate their technical choices and considerations to the jury, for example with validation metrics, test cases, research, or comparisons with reference measurements.

Which data is available?

During the hackathon we will supply you with several datasets for this challenge:

- Time-series data: hourly PM2.5 from more than 100 Hollandse Luchten sensors, some ranging several years.
- Reference measurements: Official reference station data producing hourly measurements, to possibly compare the Hollandse Luchten sensors with.
- Metadata on sensors: information about the sensor location, type, measurement communities, and intended reference stations.
- Weather data from the Dutch KNMI, for example temperature and humidity.
- Github codebase Samen Analyseren consisting of R code:
<https://github.com/rivm-syso/Analyse-Together>

Deliverables

Your submission should at least contain:

- 1) The Github repository with any used code and/or examples.
- 2) The presentation of your solution.
- 3) A concrete example showing how the system handles a question and provides an audit trail explaining how it arrived at the answer.

Every team is expected to present their solution to our jury.

Jury criteria

Will be added later

Use of tools and generative AI

AI (coding) assistants and vibe coding are allowed, as long as teams disclose clearly on where they did and did not use them. We encourage the use of such tools, but also address that teams should be able to explain and validate important parts of their solutions themselves.

Questions?

The full repository will be shared a week before the hackathon. If you have any questions, please e-mail them to janus.jansen@noord-holland.nl. These and spontaneous questions will be answered during our 'Online Q&A' on 6 November 15:00-16:00. Participants will receive the Q&A meeting link via email.