Methods of Advance Data Engineering – Data Report

How does the concentration of pollutants NO2 and O3 change throughout the day in Cuxhavener Straße, and what could be the possible reasons for it?

Data Sources

Two data sources have been used from www.govdata.de.
In order to protect human health and vegetation from the effects of excessive air pollution, air quality is continuously monitored and assessed in accordance with legal regulations.

For this purpose, the State Office for the Environment (LfU) in Schleswig-Holstein operates a network of measuring stations at which air pollutants are measured using various methods.

The measurement data from Schleswig-Holstein and a lot of additional information on the measurements are forwarded to the Federal Environment Agency and from there reported to the European Commission together with the data from all federal states.
The data links are mentioned below:

1. https://www.umweltbundesamt.de/api/air_data/v2/measures/csv?data%5B0%5D%5Bst%5D=1562&data%5B0%5D%5Bco%5D=3&data%5B0%5D%5Bsc%5D=2&date_from=2024-01-01&time_from=1&date_to=2024-12-31&time_to=24&lang=en
2. https://www.umweltbundesamt.de/api/air_data/v2/measures/csv?data%5B0%5D%5Bst%5D=1562&data%5B0%5D%5Bco%5D=5&data%5B0%5D%5Bsc%5D=2&date_from=2024-01-01&time_from=1&date_to=2024-12-31&time_to=24&lang=en

Get to Know the Data

The data sources that have been used contains data of concentration of pollutants NO2 and O3 in µg/m^3 throughout the day for every hour for the year 2024.
The data contains information about state, measuring station, time of the day when measurement has been made, the concentration of pollutant at specific time interval and the unit used.

Data Format

There are two data formats available:

1. CSV
2. JSON

Both the data formats fall under the structured data category.

Data Quality

The data from the sources fulfills good criteria to be qualified as a good quality data. Here are some metrics defined below:

- Accuracy: For this purpose, the State Office for the Environment (LfU) in Schleswig-Holstein operates a network of measuring stations at which air pollutants are measured using various methods, this ensures accuracy.
- Completeness: The data available contained 170 time-intervals with missing values of concentration out of a total of 3380 rows. The figures and values may vary as the data is being updated in real-time.
- Consistency: The units and time intervals are consistent throughout the data
- Timeliness: The data is being collected for the given time period that is for the year 2024, starting from 01.01.2024 and being collected in real-time.
- Relevancy: The data being recorded is for the required time span and for the specific area whose air quality needs to be monitored.

Data Licensing

The data is standard open-data licensed. Its terms of use are mentioned under Data License Germany Attribution 2.0, which states that
The user must ensure that the source note contains the following information:

1. the name of the provider,
2. the annotation "Data license Germany – attribution – Version 2.0" or "dl-de/by-2-0" referring to the license text available at www.govdata.de/dl-de/by-2-0, and
3. a reference to the dataset (URI).

Changes, editing, new designs or other amendments must be marked as such in the source note.

Data Pipeline

The technology being used for data pipeline are Python, pandas data frame and sqlite. The data is being read from the data sources mentioned above and then the data is being cleaned and written to the sqlite file. In the sqlite file, two tables are created one for each pollutant i.e NO2 and O3.
The data cleaning steps involves removing rows against which there was missing value

of pollutant concentration, this was done to maintain data completeness. Furthermore, a few columns were dropped because they contain data that is of no use for our analysis and pipeline. The data is continuously being updated after every hourly time interval and the implemented data pipeline takes care of it.

Result and Limitations

In the CSV files, the 3 a.m. measurement is "missing" on the day of the change from standard time (CET) to summer time (CEST), but there are two 3 a.m. measurements on the day of the change from summer time to standard time. The JSON files are not affected by this problem; standard time is used throughout.
The result of the data pipeline is a sqlite file having two tables one for each pollutant NO2 and O3. Sqlite is one of the structured data format and helps in easy understanding, can be easily queried and analyzed.
For the final report, to infer the reason of the increase in pollutants at specific time intervals is a limitation as there is no available data from the data sources, but during day times the possible reason for increase in concentration of pollutants can be traffic, and other burning of fossil fuels.