Data Report

Question

Do temperature and rainfall negatively correlate with the number of people walking through the pedestrian zone in Erlangen?

Data Sources

This analysis uses historical weather data, including temperature and rainfall measurements, along with pedestrian traffic data from yesterday to 550 days in the past. The data is collected, filtered, and prepared through Python and Jayvee pipelines before being displayed in graphs and analyzed. Since there is no public weather data source directly in Erlangen, an average between the two surrounding locations 'Möhrendorf-Kleinseebach' and 'Nürnberg Airport' is calculated to determine the weather in Erlangen.

Datasource 1: opendata.dwd.de

Structure and Quality

The data from the "Deutscher Wetterdienst" is retrieved as a ZIP file. The text file containing the data is structured like a CSV file with ";" as separator. Unfortunately, it is filled with spaces to align the content and facilitate human readability. For the automatic data pipeline this was an obstacle that had to be handled.

Data Structure (rain dataset)

- **MESS_DATUM**: Date of measurement.
- **RS**: Rainfall amount (mm).
- ...

Only attributes used for this project are listed.

Data Quality (rain dataset)

- Accuracy: High, given the structured format and clear measurement units.
- Completeness: Full dataset with no missing values in sample.
- Consistency: Consistent data points across records.
- Timeliness: Regular daily measurements.
- Relevancy: Relevant for weather analysis and trend tracking.

Data Structure (temperature dataset)

- **MESS_DATUM**: Date of measurement.
- **TMK**: Mean temperature (°C).
- ...

Only attributes used for this project are listed.

Data Quality (temperature dataset)

- Accuracy: Detailed and precise with multiple weather parameters.
- Completeness: Comprehensive dataset with no apparent missing values.
- Consistency: Uniform data collection standards.
- Timeliness: Daily observations ensuring up-to-date records.
- Relevancy: Highly relevant for detailed weather and climate studies.

Permission to use the data

See license at heading 2.2 here: https://www.dwd.de/DE/leistungen/opendata/faqs_opendata.html

Datasource 2: hystreet.com

Structure and Quality

The data from HyStreet is retrieved by an API request. The response is in JSON format.
The source provides a large amount of data, while in this project only the attributes 'timestamp' and 'pedestrians_count' were used and converted to a csv file which will be described later in the pipeline section.

Data Structure

- **Timestamp**: ISO 8601 format indicating the date and time.
- **Pedestrians Count**: Integer for the total number of pedestrians.
- ...

Only attributes used for this project are listed.

Data Quality

- Accuracy: Data appears accurate and detailed, with specific counts for various categories. Potentially verified by the *unverified* flag.
- Completeness: Comprehensive data with detailed breakdowns. The *unverified* flag indicates potential gaps or unconfirmed data.
- Consistency: Structured and consistent format across entries and consistent use of temperature and pedestrian count units.
- Timeliness: Recent data with timestamps indicating when the data was recorded.
- Relevancy: Detailed information relevant to pedestrian and weather analysis. Breakdown by zones and demographic categories adds context and specificity.

The HyStreet data source provides well-structured and detailed data, suitable for easy querying and in-depth analysis.

Permission to use the data

See permission to use the data at heading 4 here: https://hystreet.com/agb
Additionally the permission was granted per email, to use the data from the last 3 years of the location 'Erlangen'.
For grading: If a proof of the email conversation is needed, please contact me.

Data Pipeline

Pipeline 1

- API Call
- Timezone
- Filter JSON
- Convert JSON to CSV

Pipeline 2

- Load CSV
- Retrieve 4 ZIP files
- Manually changing file names (dates in it are changing)
- put everything into sqlite database

Pipeline 3

- Calculating rain and temperature averages for all entries of the two weather data locations
- Write the prepared data into new table

Result and Limitations

As this report is limited to 3 A4 pages, I am referring to the plots in the 'plots' folder instead of adding them to the report.

When analyzing the data, there are some outliers in the number of pedestrians. This is due to some special events that take place in Erlangen every year.

This is the list of the events for the data between November 2022 and June 2024:

- 2022-11-21 - 2022-12-23: "Weihnachtsmarkt"
- 2023-05-25 - 2023-06-05: "Bergkirchweih"
- 2023-10-21: "Lange Nacht der Wissenschaft"
- 2023-11-25 - 2023-12-23: "Weihnachtsmarkt"
- 2024-05-16 - 2024-05-27: "Bergkirchweih"
- 2024-05-30 - 2024-06-02: "Internationaler Comic Salon"