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

The pipeline script "pipeline.sh" deletes the database, if it already exists and triggers three sub-pipelines, which are described below.

Pipeline 1 ("pipeline1.py")

The first pipeline handles the API call to retrieve the data from HySteet. To get the data for a specific time period, a "from" and "to" date must be specified. In addition, this date must be in UTC format. The "to" date is calculated as the last second of the previous day and the "from" date is the first second of the 550th day before the "to" date. The API request returns the requested data in JSON format. Because not all attributes are needed for this project, the returned JSON is filtered. The JSON data is then converted to CSV format and exported to a file that is accessed in the next sub-pipeline.
The filenames from the weather data source are changing every day at around 11:30 am CET (when the new data is being released). In order to make the pipeline work automatically, the filename strings are being adapted by the second part of this pipeline. **Caution:** Between 11:25 and 11:35 it is likely that the pipeline crashes, depending on the exact release time of the updated weather data.
For this pipeline, Python was used because it provides many useful libraries for the challenges described above.

Pipeline 2 ("pipeline2.jv")

The second pipeline is written in Jayvee. It loads the CSV file with the pedestrian data and downloads the four ZIP files with the rain and temperature data from the provider's website. The datasets are all from the previous day to the 550 days in the past. That's why these dates were calculated in the first pipeline for the API call. The different datasets can then be easily merged in the third pipeline. After filtering for the relevant columns, the five datasets are loaded into a SQLite database.

Pipeline 3 ("pipeline3.py")

The third pipeline again uses Python to execute SQL statements in the database created earlier. In order to have all the necessary data in one place for easier analysis, a new empty table is created in the database. As mentioned before, the average of the two rain and temperature datasets needs to be calculated. This was done using temporary views, merging all together with the pedestrian data and inserting everything into the new empty table. The data can now be easily analyzed.

Result and Limitations

Output data of the pipeline

The output data of the pipeline is a single SQLite database, with a table that combines pedestrian traffic data with averaged weather data (temperature and rainfall) for the corresponding days. The data spans from yesterday to 550 days in the past, providing a comprehensive dataset for analysis.

Data Structure

- Date: Date of the measurement.
- Pedestrian Count: Number of people walking through the pedestrian zone in Erlangen.
- Average Temperature: Mean temperature in Erlangen (°C), averaged from the surrounding locations.
- Average Rainfall: Rainfall amount in Erlangen (mm), averaged from the surrounding locations.

Data Quality

- Accuracy: High, given the structured and detailed nature of the source data.
- Completeness: The dataset is complete with no missing values in the sample provided.
- Consistency: Data points are consistently recorded across all days in the dataset.
- Timeliness: Regular daily measurements ensure up-to-date records.
- Relevancy: High, for analyzing the relationship between weather conditions and pedestrian traffic.

Reflection on the data

The data is comprehensive and relevant for the analysis of the correlation between weather conditions and pedestrian traffic. However, there are a few potential limitations:

- Temporal Resolution: Daily aggregates of weather data and pedestrian counts might obscure more granular patterns (e.g., hourly variations).
- Spatial Averaging: Averaging weather data from two locations may not perfectly represent the actual weather in Erlangen, potentially introducing some bias.
- Data Update Timing: The necessity to adjust the hardcoded filenames in the pipeline to accommodate new data releases can introduce delays or errors if not managed correctly.

Potential issues for the final report

- Data Integration: Ensuring seamless integration and synchronization of data from different sources can be challenging.
- Data Averaging: The averaging method for weather data from surrounding locations could be questioned for its accuracy in reflecting actual conditions in Erlangen.
- Pipeline Robustness: The need for manual adjustments in the pipeline could affect its robustness and reproducibility. Automating this process or finding a workaround would be beneficial.
- Analysis Limitations: The correlation analysis may be limited by the data resolution and potential biases introduced by the averaging process. Addressing these limitations in the report will be crucial for transparency and reliability of the findings.

Early insight to the results

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:

- 2023-05-25 - 2023-06-05: “Bergkirchweih”
- 2023-10-21 - 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”

As this report should not exceed three A4 pages, I am only referring to the plots in the folder “plots”.

Since these plots are saved as vector graphics, you can zoom in to see all the details.