Analysis Report

July 3, 2024

Introduction

This analysis report, aims to investigate the correlation between temperature and rainfall with the number of people walking through the pedestrian zone in Erlangen. The main question is the following:
"Do temperature (positively) and rainfall (negatively) correlate with the number of people walking through the pedestrian zone in Erlangen?"
In addition it would be interesting which other factors influence the number of pedestrians in Erlangen. By examining the data and conducting relevant analyses, we can gain insights into the relationship between these weather factors and pedestrian behavior.

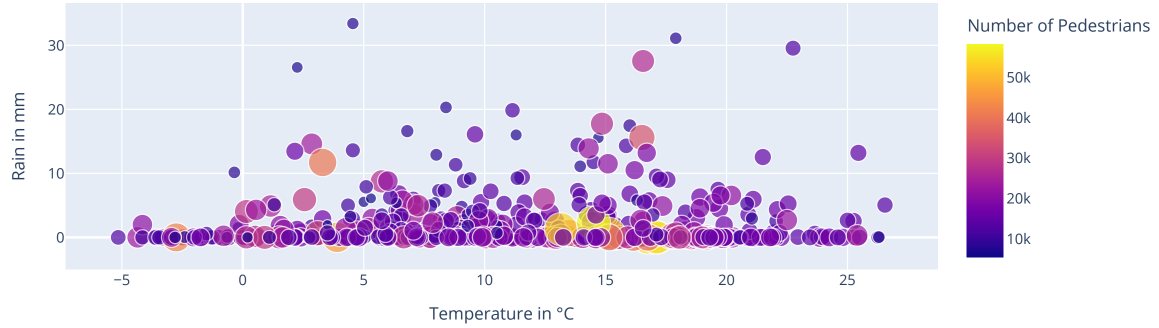
Used Data

The data used for the analysis is the output of the data pipeline. When the data is up to date (can be achieved by running `pipeline.sh`) it represents the number of pedestrians together with the average temperature and rainfall per day starting from yesterday up to 550 before that day. Data from the company "HyStreet" and the "Deutscher Wetterdienst" have been combined into a single table of an SQLite database.

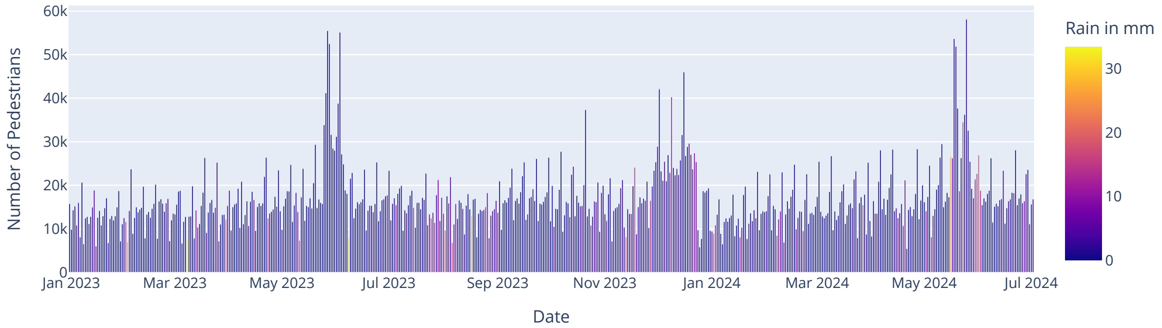
At this point I would like to express my gratitude to the company HyStreet for providing me with historical data of the pedestrian zone in Erlangen.

Analysis

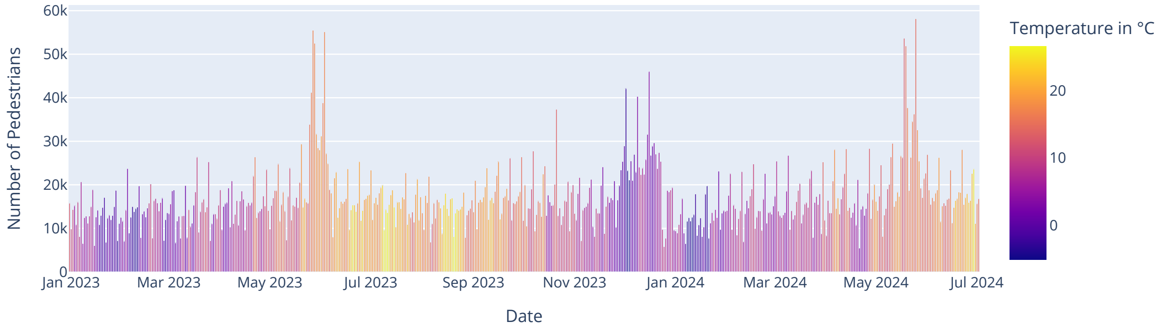
Before analysis, the data has been explored with a Jupyter Notebook (`data-exploration.ipynb`). The following plots have been created on the latest project update (*date below title of this document*):



Rainfall and temperature in relation to the number of pedestrians.



Number of pedestrians per day colored with rainfall in mm.



Number of pedestrians per day colored with temperature in °C.

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

To be able to answer the question from the introduction, the data needs to be *cleaned up* to prevent the mean values to be distorted. For example the number of pedestrians is strongly depending on the weekday. Usually on Saturdays the number is highest and on Sundays lowest of the week. That's why each day should only be compared to the same weekdays while considering the weather data. In addition there are a few events taking place in Erlangen over the year, where a lot more pedestrians are in the streets. These events need to be filtered out before the analysis.
The following steps have been implemented in the file `analysis.py`:

1. Remove the days that are within an event that takes place in Erlangen (e.g. Bergkirchweih, Weihnachtsmarkt, . . .)
2. Calculate the average number of pedestrians per weekday.
3. Select the days with more than 10% difference to the median of that weekday.
4. For the remaining days, calculate the correlation between pedestrian counts and temperature and rainfall, separating the data with pedestrian counts above and below the median.

The following code block contains the output of the Python analysis script:

```
This script has been executed on the 2024-07-03, with daily data from the
period 2022-12-31 to 2024-07-02.

Median number of pedestrians per weekday (without the event periods):
Monday:    14149
Tuesday:   15191
Wednesday: 14571
Thursday:  15309
Friday:    17398
Saturday:  22971
Sunday:     9753
```

```
Statistics for days when the number of pedestrians differs by more than
10% from the median for that weekday.

Days with more than 10 % above the median:
Average temperature: 15.06 °C
Average rainfall:    0.58 mm

Days with more than 10 % below the median:
Average temperature: 6.93 °C
Average rainfall:    4.16 mm
```

Conclusions

In order to be able to answer the original question, the data was cleaned by filtering out days with events in Erlangen. For the remaining days the median for each weekday was calculated and days with a difference of more than 10 % to the median were selected. The average temperature and rainfall for the days with a pedestrian count above and below the weekday median were calculated.

With these adjustments, the question from the introduction can be answered:
> **Yes, temperature is positively and rainfall negatively correlated with the number of pedestrians in Erlangen.**

While analyzing the data another conclusion was drawn:
> **Social events and Sundays have more influence on the number of pedestrians in Erlangen than rain or temperature**.

Limitations

The correlation analysis may be limited by the data resolution as the weather data used is averaged (temperature) / summed (rainfall) per day. For example, if it rains a lot at night, when there are usually fewer pedestrians on the street, and it does not rain during the day, the analysis will be distorted. In addition, there may be other factors that influence the number of pedestrians, such as school holidays or major events in the proximity of Erlangen. These factors were not taken into account in this analysis.