Project Report

FAU Erlangen-Nürnberg

Professorship for Open-Source Software

Course: Advanced Software Methods / Software Applications with AI

Project Title: Analyzing Bicycle Traffic in the City of Münster (Germany)

Course Instructors: Philip Heltweg & Georg Schwarz

Student: Martin Reimer

Submission Date: 27.06.2023

Table of Contents

1. Introduction

The analysis of bicycle traffic patterns holds significant importance for urban planning, as it provides crucial insights into the utilization of cycling infrastructure and the behavior of cyclists. Understanding these patterns allows for the identification of areas where bicycle infrastructure is needed and the optimization of existing infrastructure to better serve the needs of cyclists.

This data science project aims to explore the trends and patterns of bicycle traffic at specific stations in **Münster, Germany**, over a given time period. By analyzing factors such as time of day, external influences like weather conditions and holidays, we can gain valuable insights into the dynamics of bicycle traffic in the city.

The project relies on two primary data sources:

1. **Verkehrszählung Fahrradverkehr: Tagesaktuelle Daten - Münster:** This dataset provides daily counts of cyclists at various bicycle counting stations in Münster. By examining this data, we can identify peak traffic periods, popular cycling routes, and potential areas for infrastructure improvement.

- mobilithek URL
- Data URL
- Official Documentation
- Data Type: CSV
- License: freie Nutzung / Open Data

2. **meteostat.net:** Weather conditions play a crucial role in influencing cycling behavior. The meteostat.net dataset offers access to weather data from national meteorological offices, allowing us to analyze the correlation between weather conditions and bicycle traffic patterns. This information can guide urban planners in developing weather-resilient cycling infrastructure and promoting cycling as a sustainable transportation option.

- Official Documentation
- Data Type: CSV
- License: CC BY-NC 4.0

3. **Feiertage-API:** Public holidays can have a significant impact on bicycle traffic patterns. The Feiertage-API provides access to public holiday information, allowing us to analyze the relationship between holidays and bicycle traffic volume. This information can provide insights into how holidays influence cycling behavior.

- Documentation
- Beispiel Abfrage
- License: no license / free of charge

- Data Type: JSON

The utilization of these data sources provides a comprehensive understanding of bicycle traffic patterns in Münster, facilitating informed decision-making for urban planners and policymakers. By analyzing the available data, we aim to uncover valuable insights that can contribute to safer, more accessible, and efficient cycling infrastructure in the city.

Through this project, we seek to demonstrate the significance of data-driven analysis in improving urban planning and transportation systems. By leveraging the power of data science techniques, we can gain actionable insights that have the potential to reduce traffic congestion, improve air quality, and enhance the overall cycling experience in Münster.

2. Project Implementation

In the implementation phase of the project, an initial and incremental data pipeline was developed to efficiently handle the dataset updates. The decision to create two separate pipelines was motivated by the need to avoid downloading the entire dataset every day, which would not be efficient.

The initial data pipeline is responsible for retrieving the data from the GitHub repository and extracting the metadata and file data for all counters. It generates a CSV file name based on the provided datetime and retrieves the file contents accordingly. This pipeline ensures that the most recent data is always obtained.

On the other hand, the incremental data pipeline addresses the need to append new data to the existing dataset. It retrieves the incremental data for the current month and appends it to the dataset, keeping it up to date without unnecessary duplication of data. This approach significantly reduces the computational resources required for data processing.

To facilitate code management and maintainability, the ETL process was split into three files:

- Extractor: Provides methods to extract data from each source, including the bicycle traffic data, weather data, and holidays data. It retrieves the relevant information and organizes it into a dictionary format for further processing.
- Transformer: Transforms the extracted data into a suitable format for analysis. It converts the extracted data into pandas DataFrames, adjusts datetime values to the appropriate timezone, merges the transformed DataFrames from different sources, applies forward-fill to handle missing values, creates an **is_holiday** column based on holidays presence, and removes unnecessary columns. The transformed DataFrame serves as the basis for data exploration and analysis.
- Loader: Establishes a connection to a SQLite database and loads the transformed data into it. The `load_initial_data` function replaces the entire dataset, while the `load_incremental_data` function appends new rows to the existing dataset based on the maximum datetime value. Additionally, both functions save the loaded data as CSV files for reference and provide a reliable and efficient storage solution.

Testing and Continuous Integration (CI) were essential components of the project implementation. The test.sh script was created to execute two ETL tests: one for the initial data pipeline and another for the incremental pipeline. The script checks the existence of specific output files to determine the success or failure of each test.

By integrating the `test.sh` script into the CI pipeline using GitHub Actions, automated testing was performed on every push. This allowed for the execution of the `test.sh` script and the validation of the output files. The continuous integration process helped identify any issues or inconsistencies in the data pipeline, ensuring that the project maintained a high level of quality throughout its development.

3. Data Results & Discussion

3.1 Data Summarization and Missing Data

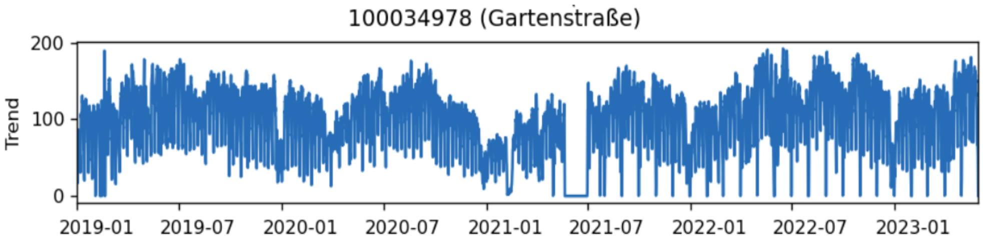
To simplify data visualization and analysis, the counters of each street were summed up into one counter. This approach facilitated the interpretation of overall traffic trends and eliminated the need to analyze each sub-counter separately. However, if any of the sub-counters for a street were unavailable on a particular day, the entire dataset for that street counter was set to NaN. This ensured data integrity but introduced challenges in interpreting the results.

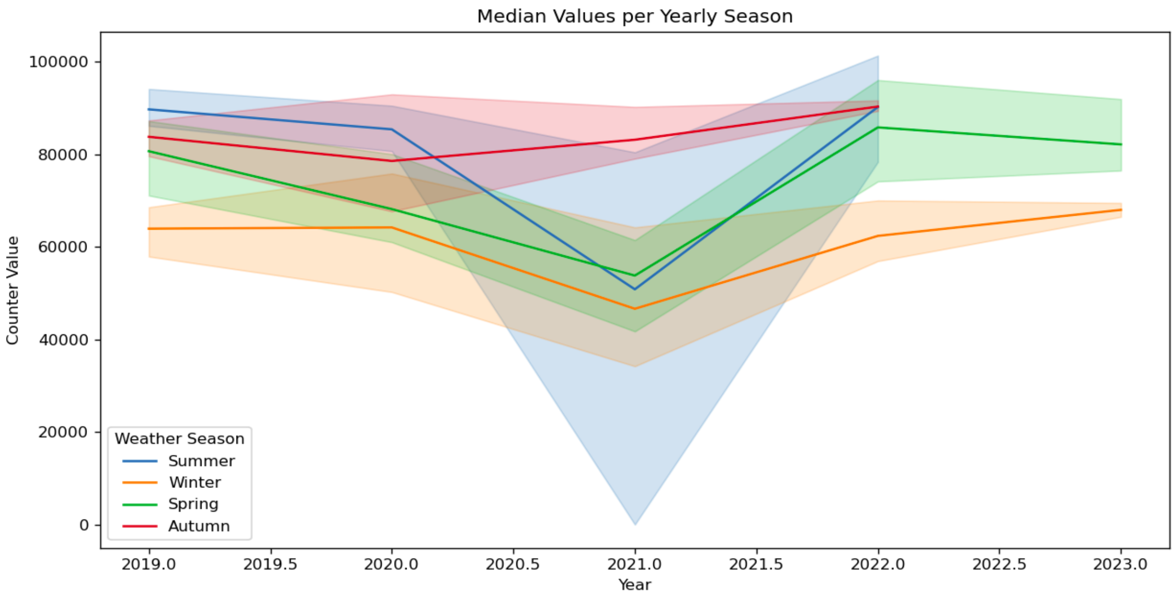
3.2 Categorical Analysis

Several categorical analyses were performed to gain insights into bicycle traffic patterns:

3.2.1 Seasonal Decomposition

The seasonal decomposition analysis revealed expected patterns in bicycle traffic, as is shown in the trend analysis of the Gartenstraße (Garten street). This specific counter exhibited low traffic during winter, followed by a significant increase in spring and a peak in summer. These trends align with general expectations for cycling behavior.

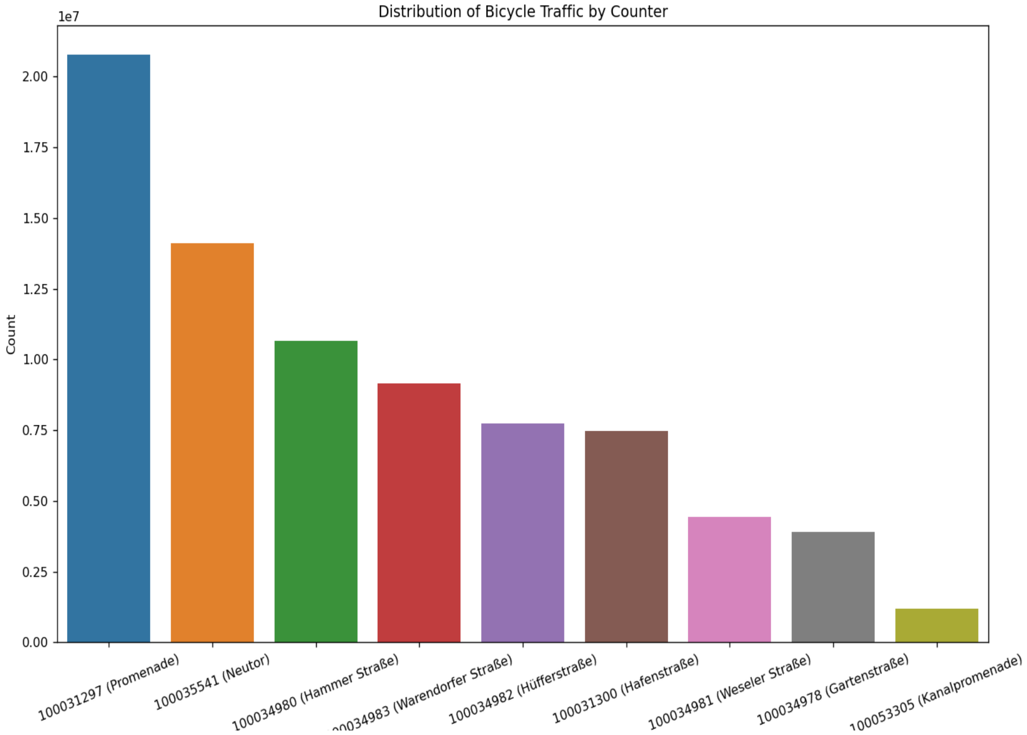




The Median Counter per Season graph displays the yearly development of the seasonal counters, providing insights into season-specific trends. However, it is noteworthy that except for autumn, all seasons had relatively low counter values in 2021, which may be due to missing data or construction work. Further investigation is needed to determine the underlying reasons for this unexpected pattern. The analysis over the three-year period does not indicate a clear increasing or decreasing trend in bicycle traffic.

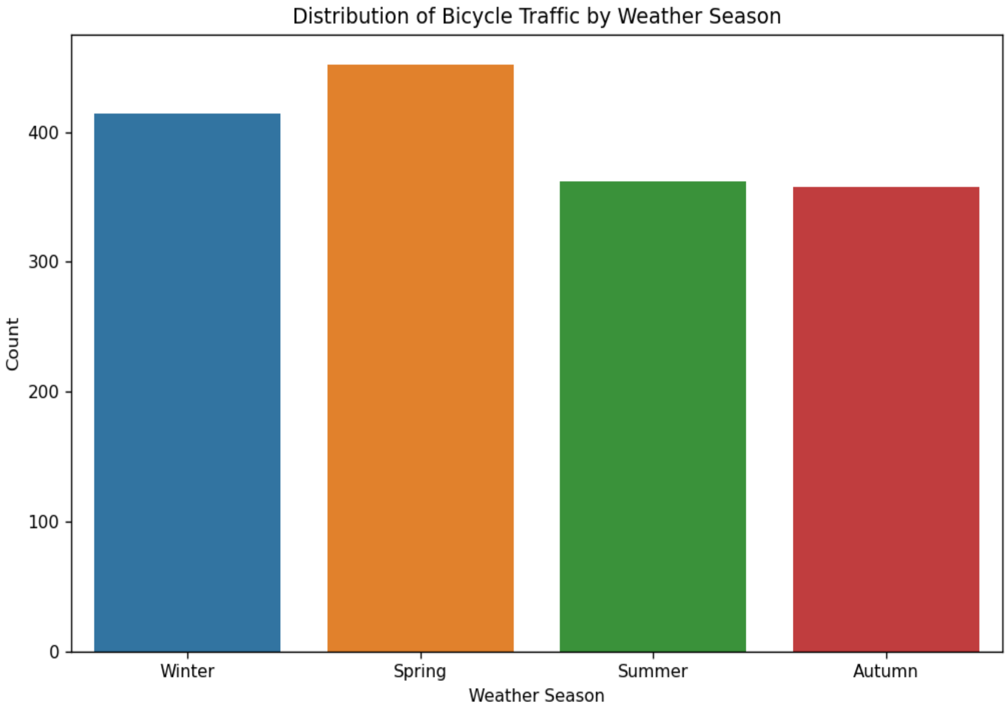
3.2.2 Distribution of Traffic

The distribution of bicycle traffic among counters showed significant variations. For instance, the Promenade counter recorded the highest count, with approximately 20,000,000 bicycles from 2019 to May 2023. On the other hand, the Kanalpromenade counter had the lowest count of around 1,100,000 bicycles. This discrepancy highlights the unequal distribution of bicycle traffic across different counters.



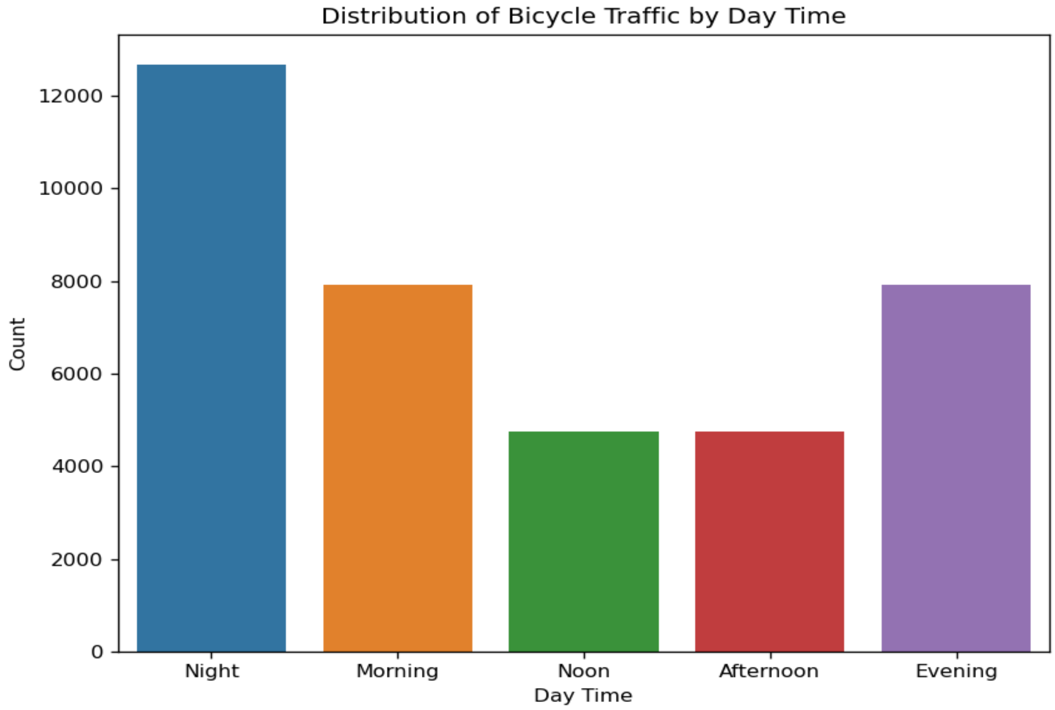
3.2.3 Distribution by Weather Season

The analysis of bicycle traffic by weather season revealed that spring had the highest number of bicycle riders, followed by winter. The unexpected placement of winter before summer in terms of traffic could be attributed to missing data problems encountered or too hot weather during summer season. Further investigation is needed to determine the underlying reasons for this unexpected pattern.



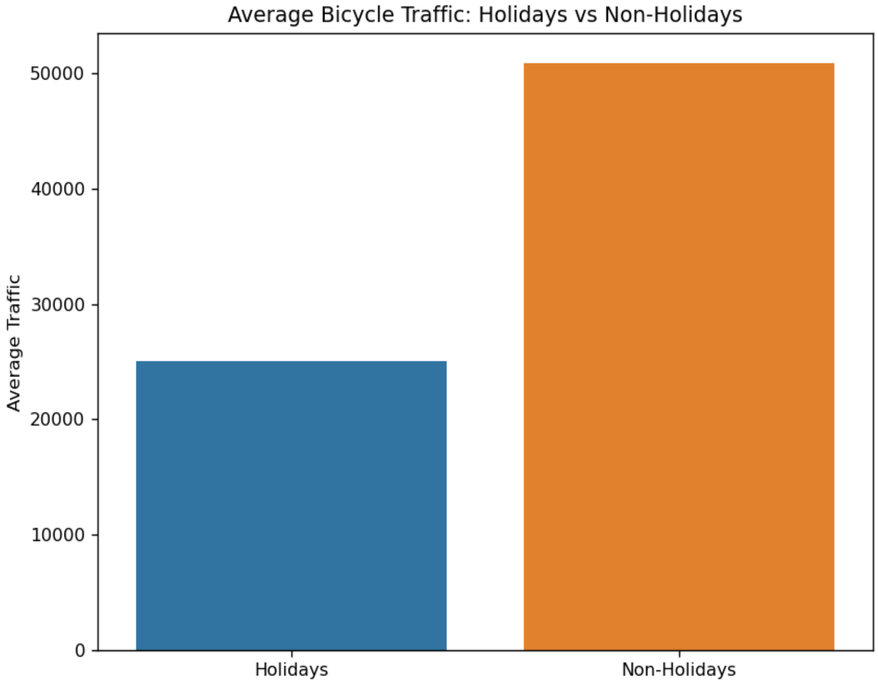
3.2.4 Distribution by Day Time

Surprisingly, the night was identified as the daytime period with the highest total traffic. This finding may be influenced by the definition of night, which spans eight hours, while other periods such as noon are defined within only three hours. Additional factors contributing to this pattern should be explored. For example, it is possible that under certain circumstances, some counters do not store hourly but daily data, and this is represented as the count at 0:00, which corresponds to night. This hypothesis should be further investigated in subsequent analysis.



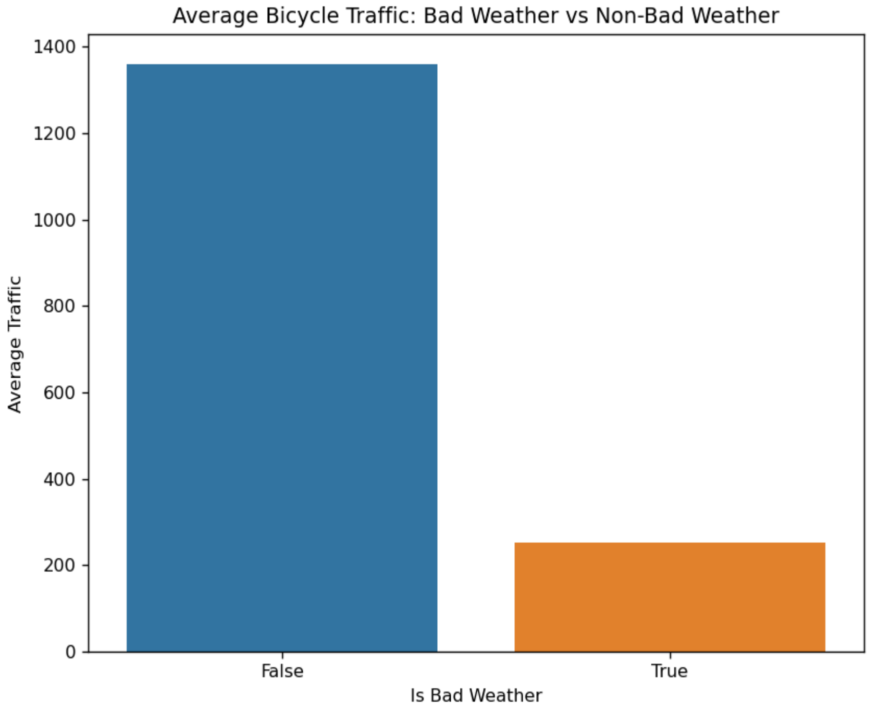
3.2.5 Holiday vs. Non-holiday

On average, bicycle traffic on holidays was approximately half of that on non-holidays. Further analysis could focus on exploring the impact of individual holidays and investigating whether certain holidays experience increased traffic.



3.2.6 Bad Weather vs. Non-bad Weather

In this analysis, a bad weather day was defined as a day with precipitation higher than 0.1. It is important to note that rain is just one factor influencing the subjective perception of good or bad weather. Nonetheless, the analysis based on this definition reveals that rainy days were associated with significantly lower bicycle traffic. This finding underscores the influence of weather conditions on cycling patterns. Further analysis should delve deeper into whether this finding demonstrates correlation or causality. One argument could be that rainy days mostly occur during winter in Germany, and rain might not be the sole reason for the observed decrease in bicycle traffic.



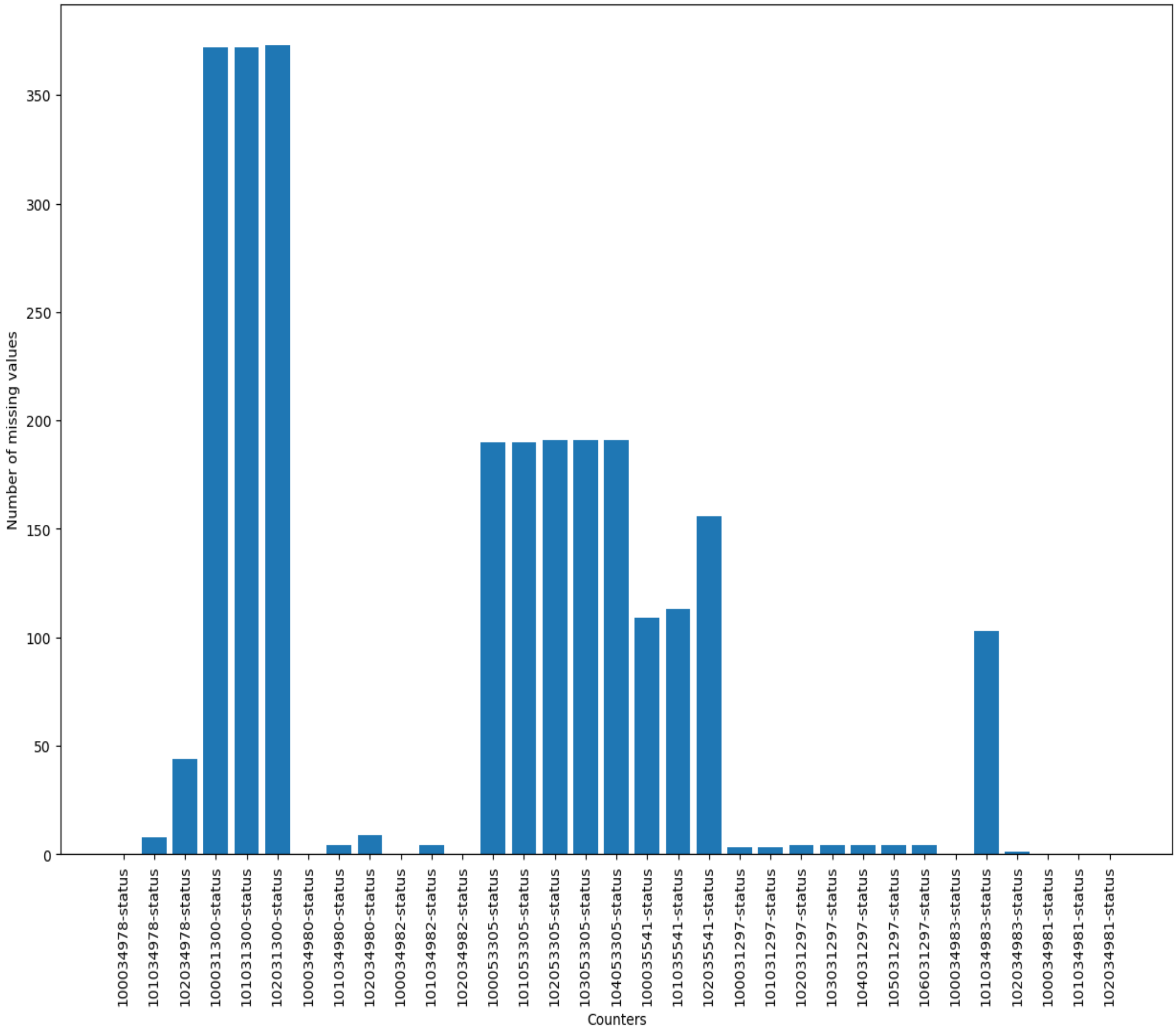
3.3 Dataset Challenges

The analysis encountered several challenges related to the dataset:

3.3.1 Missing Data

The dataset presented significant challenges due to the presence of missing data resulting from maintenance or construction work on the streets. To ensure data integrity, missing values were not filled in, as doing so would introduce biases into the analysis. The extent of missing data varied across different counters, with some experiencing sporadic missing values for only a few days, while others had missing values for over 30 days multiple times throughout the three-year period. These missing data points complicated visualizations and made the analysis more challenging, as the presence of construction work had to be taken into account when interpreting the results. However, despite these challenges, the analysis still provides valuable insights into bicycle traffic patterns.

The figure below illustrates the total number of missing values in the dataset. It is important to note that some counters were set up much later than others, which is why they have a considerably higher number of missing values. However, nearly half of all counters have missing data for at least 100 days, which poses additional difficulties for analyzing the dataset over a three-year basis.



3.3.2 Missing Weather Data

The weather data obtained from the Meteostat API had limitations. The dataset only included weather data from July 14, 2021, onward, leaving a gap of approximately 2.5 years without weather information. Furthermore, even after this date, several weather columns were missing for extended periods. To address this issue, the analysis focused on three selected weather features, which had continuous values: temperature, precipitation, and wind speed.

3.4 Limitations

Due to the scope and constraints of the project work, the analysis may not have delved into every aspect or variable that could potentially influence bicycle traffic patterns in Münster. While the analysis provided valuable insights into certain factors such as time of day, day of the week, weather conditions, and holidays, it may not have explored all possible variables that could contribute to the observed patterns. A more in-depth analysis could involve considering additional factors such as infrastructure quality, population density, or socioeconomic factors, which may provide a more comprehensive understanding of bicycle traffic patterns.

4. Conclusion

In conclusion, this project represents a valuable undertaking in the sphere of urban planning, specifically concerning the cycling infrastructure in Münster, Germany. Through the application of data science methodologies, we analyzed bicycle traffic patterns and investigated the influence of various factors such as time of day, weather conditions, and holidays.

Our data sources, which included **Verkehrszählung Fahrradverkehr: Tagesaktuelle Daten**, **meteostat.net**, and **Feiertage-API**, were instrumental in painting a comprehensive picture of bicycle traffic patterns in the city. We utilized an efficient data pipeline, separating the process into initial and incremental phases to optimize resource usage and maintain data integrity.

The study uncovered an unequal distribution of bicycle traffic across different counters, with the Promenade counter recording the highest counts. Interestingly, the data indicated a higher volume of traffic during night hours, which warrants further investigation for underlying causes.

Moreover, holidays and weather conditions, especially rainy days, were found to influence cycling behavior, with traffic volumes being substantially lower during these periods. The impact of weather conditions underlines the importance of weather-resilient infrastructure in promoting cycling as a viable means of transportation.

However, the analysis was not without challenges. We encountered missing data in both the bicycle counters and weather datasets, which introduced complications in visualization and interpretation. Furthermore, due to the scope and time constraints of the project, the analysis was not exhaustive in considering all possible factors influencing bicycle traffic patterns.

Overall, this project contributes valuable insights into bicycle traffic patterns in Münster and emphasizes the importance of data-driven analysis in urban planning. The findings can inform decision-making processes to improve cycling infrastructure and promote sustainable transportation in the city.