TIL6022 TIL Python Programming Final Report

Correlation between transport modes and weather conditions

Group 15

Name	Github username
Brodbelt Lopez, Nicolas David	NBrodbelt
Hemalatha Thirumal, Manasa	Manasaht31
van Herwijnen, Anna	AnnavanHerwijnen
Nikolova, Greta	greta.sch4-commits
Zeng, Zitong	Zitong22

Author contribution statement:

- Anna – Finding relevant/usable datasets, gaining access to protected datasets, creating and updating main research question/goal, downloading and processing of ovbike data and knmi data, correlation plot creation for ovbike and knmi data
- Evelyn – introduction writing and explanation of the transport modes, correlation between translink and temperature & rainfall, creating the plots of translink dataset , writing the introduction
- Greta – Understanding the structure of the databases, description of the practical information correlated to them, group discussions, part of the preprocessing of OV-fiets data, writing and structure of the Jupyter Notebook, review of the final draft and redacting inconsistencies, adding explanation comments to the code plotting the Translink linear regression
- Manasa – Making a Data Pipeline for the data visualization after going through the dataset and various other data pipeline architecture, working on the correlation between the weather parameters throughout the year initially, the KNMI dataset, and report writing and phrasing
- Nicolas – Research question formulation, exploring possible datasets, finding prior research on the topic for hypothesis formulation, hypothesis formulation, writing final version of section 1,2 and 3, rewriting section 4,5,6,7, writing conclusion and initial discussion

points.

Introduction

Every day, the daily habits of countless individuals are altered by the weather they encounter. Studies have shown that weather significantly affects mode choice; dry, calm, and moderately warm conditions encourage cycling. Likewise, adverse weather reduces cycling and increases public transport use (Böcker et al., 2016). Additionally, research on e-bike travel indicates that rain, wind, and snow reduce usage, while moderate temperatures have a positive effect, very high temperatures may decrease trips. Lastly, concerning public transport, adverse weather can increase travel time and affect user willingness (Sabir et al., 2010).

However, most research focuses on a single mode or on local areas, with limited studies examining nationwide weather variations across several transport modes. As such, it can be said that it remains unclear how weather drives transport mode shifts at a national level in the Netherlands' transport system.

Understanding these dynamics is crucial for climate adaptation in the years to come. By gaining greater understanding on this topic, operators can optimise scheduling and resource use. Likewise, policymakers can design more resilient sustainable transport networks that account for such patterns.

This project aims to examine the relationship between daily weather patterns and travel mode choice in the Netherlands from 2023 to 2024. This is done by means of the use of weather data (from KNMI), train check-ins (Translink) and shared bike (OV Fiets) availability. By means of such data, an analysis has been conducted to evaluate transport preferences within cycling and train ridership. Specifically, by using weekly averages for train check-ins and OV Fiets usage, this could be compared to the maximum weekly temperature and total weekly rainfall.

Research Questions

Based on the research aim defined above, the following research question has been defined:

What is the correlation between daily transport (train and bicycle) usage and weather patterns (average temperature and precipitation) in urban areas in the Netherlands in 2023?

In order to operationalise this research question within the context of the TIL6022

project, more manageable research sub-questions have been defined:

- What is the correlation between train ridership and weather patterns (average temperature and precipitation)?
- What is the correlation between shared bicycle (OV Fiets) usage and weather patterns (average temperature and precipitation)?
- How does train ridership shift with changes in weather conditions (average temperature and precipitation)?
- How does shared bicycle usage shift with changes in weather conditions (average temperature and precipitation)?
- How does the difference caused by different weather conditions compare between train ridership and shared bicycle usage?

It is important to note that the weather patterns that can be analysed are dependent on data availability. This is particularly important with respect to location-specific weather patterns, such as wind or precipitation levels.

Databases Used

Based on the research questions and sub-questions formulated, numerous possible databases were explored and evaluated for possible use. Of the databases found, three were subsequently used:

- The KNMI Weather Database
- The Translink Public Transport Database
- The OV Fiets Availability Database.

These databases have been further detailed below:

1. Koninklijk Nederlands Meteorologisch Instituut (KNMI, or the Royal Dutch Meteorological Institute) Weather Dataset:

Firstly, the KNMI database provides comprehensive meteorological observations collected from multiple weather stations distributed throughout the Netherlands. It includes parameters such as daily minimum and maximum temperature, wind speed, wind gusts, sunshine duration, and precipitation duration. Each station’s geographic coordinates enable spatial linking with nearby public transport stations. The historical data spans from 1991 to 2025, ensuring a robust temporal coverage. From the wide range of data available, only a few of these were found to be sufficiently available at a geographical level to allow for subsequent analysis.

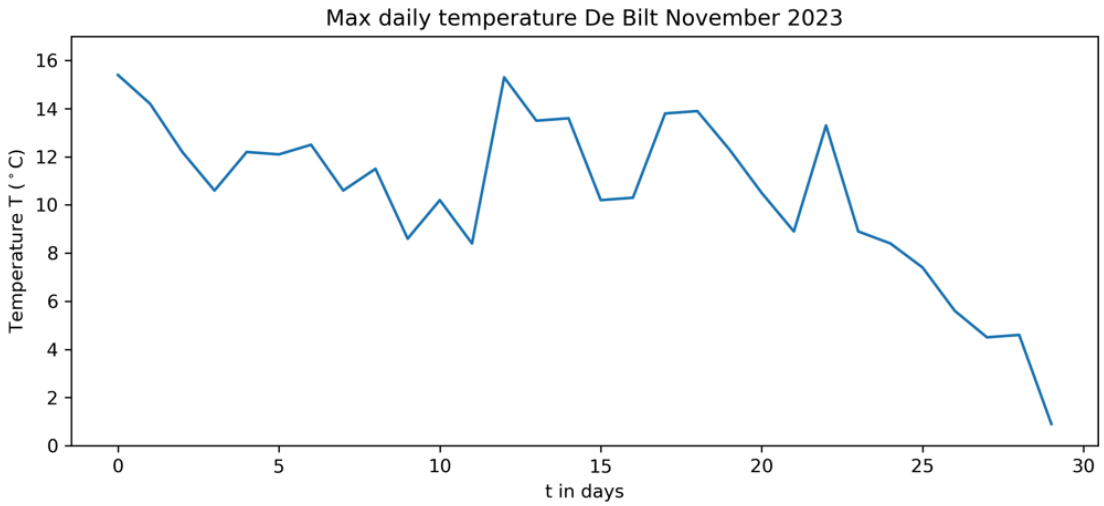

Figure 1 - Maximum daily temperature in De Bilt, November 2023

As seen in Figure 1, the precipitation and temperature for any given station are recorded in hourly intervals. Of these, it is important to note that the temperature data is taken solely from De Bilt (which acts as the national average) (KNMI, n.d.)

Below, in Figure 2 the geographical location of each weather station within the KNMI weather database can be observed. This has then been juxtaposed with the weather stations that were used for subsequent analysis. These were chosen by writing code to select the closest station to the OV Fiets rental locations that are of interest for the project.

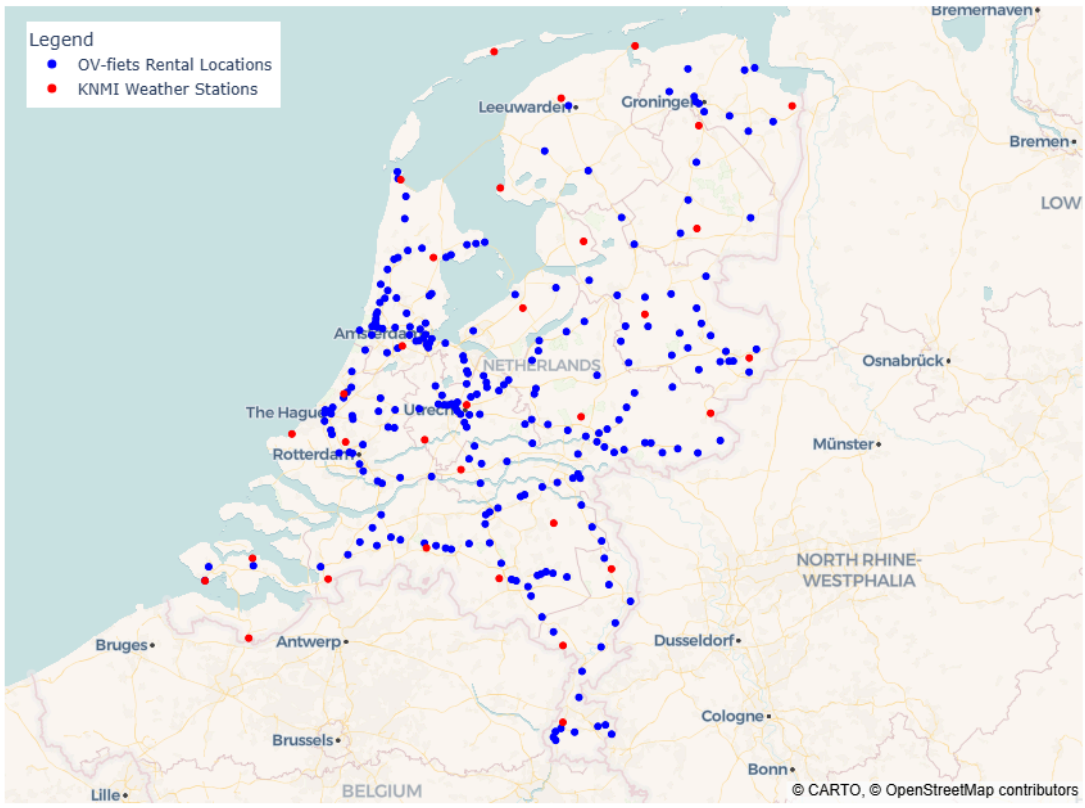

Figure 2 - Geographical locations of weather stations and OV-fiets locations

2. TransLink Public Transport Database

Then, the TransLink database contains aggregated information on train check-ins within NS (the main rail operator in the Netherlands). It accounts for total check-ins using either an OV Chipkaart, or using OVpay (that was introduced in 2023). The database provides data from 2019 until October of 2025. It is important to mention that the database only considers national check-in data and does not include station-specific check-in data.

As seen in Figure 3, the total number of check-ins for each day in November 2023 is recorded. Most notably, the plot shows the cyclical weekly nature of check-in's, with weekdays having notably higher check-in's than weekends.

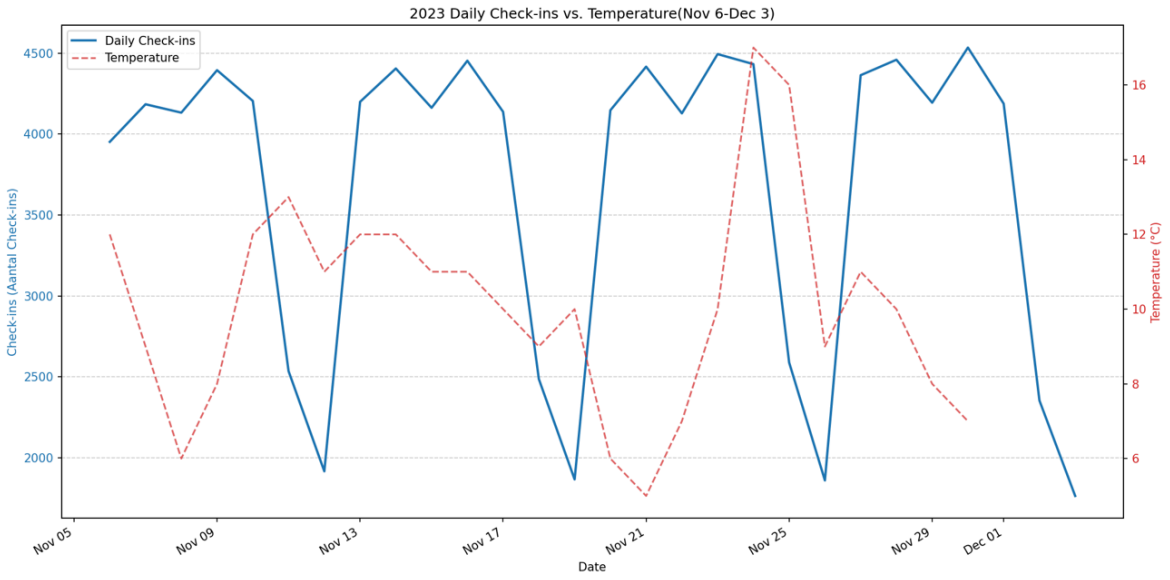

Figure 3 - Daily train check-ins for November 2023

3. OV Fiets Availability Dataset

Lastly, the OV-fiets database is an open-source historical archive derived from the NS API, documenting the availability of rental bicycles at public transport stations. Each record includes the timestamp of data logging, the number of bikes available, and the station location. The database covers the period from 2015 to early 2025, offering a long-term perspective on shared bike usage patterns.

It is important to note that this data is obtained by means of an un-official database created by Adriaan van Natijne (n.d.) using data from NS's API. As such, several workarounds were necessary in order to use such data. Specifically, the way the database is updated is by periodically checking every 15 minutes how many bikes are available at each station.

Below in Figure 4, a visualisation can be seen based on the daily availability for station Delft.

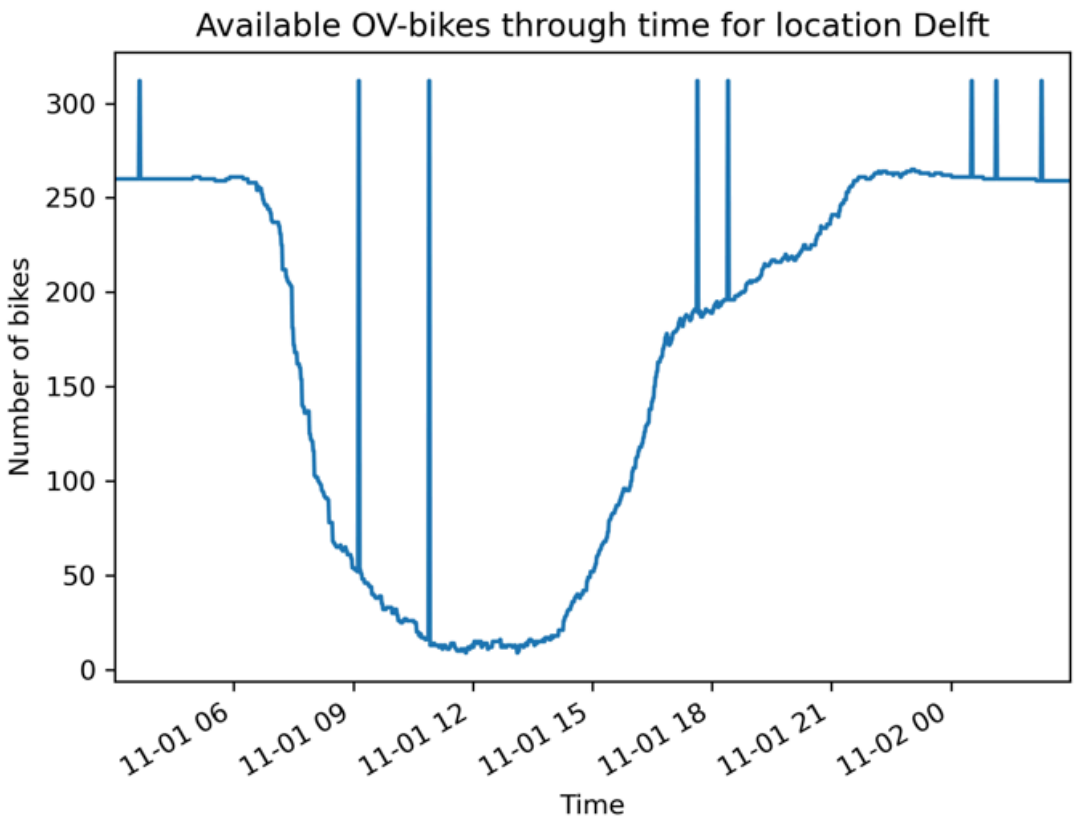

Figure 4 - Available OV-bikes in Delft

Links to databases:

- KNMI weather dataset: KNMI - https://www.knmi.nl/nederland-nu/klimatologie/daggegevens
- TransLink dataset: https://translink.nl/open-data/
- OV-fiets availability dataset: https://trein.fwrite.org/idx/dedup_OVFiets.html

Research Gap and Hypothesis Formulation

Based on the research question posed, it is now important to briefly explore the prior art. This serves a dual purpose, on the one hand, such an effort helps find what the current project can contribute to. On the other hand, this process helps identify what relations could be expected between the different variables. This has been done by exploring outcomes reached by prior research and subsequently complemented by a newly formulated hypothesis.

It was found that prior research had already been conducted in the desired

geographical context. Namely, investigations by Böcker et al. (2019) and Wilkesmann et al. (2023) were conducted within the desired geographical context and focused on the effect of weather on mobility patterns. More concretely, Böcker et al. (2019) analysed the effect of weather on mode choice and Wilkesmann et al. (2023) evaluated possible “temporal and weather-related determinants” associated to the use of OV Fiets. Beyond this, research by Galich and Nieland (2023) was also found to be insightful, covering the effect of weather conditions on the transportation mode used in Germany.

From the literature studied, the key research gap reached is related to the analyses conducted with seasonal data and more detailed analyses with different weather phenomena such as wind, precipitation or fog. Therefore, efforts will be made to assess the feasibility of conducting such research with the databases used in the given context of the TIL6022 project.

Then, regarding the hypothesis formulation concerning the effect of weather on transport mode selection, the research by the papers mentioned attribute a varying degree of significance to the effect weather has on transportation mode selection. For instance, Galich and Nieland (2023) characterise this effect as marginal – most notably concluding that these impacts tend to be smaller in densely-populated urban metropolis – while Böcker et al. (2019) find considerable regional differences in this effect. Lastly, Wilkesmann et al. (2023) state that no clear set of variables were found that were able to explain variance across the entire set of stations. Due to the differences in findings amongst the papers mentioned, it was not possible to create a hypothesis linked to the general trend of research.

As such, a new hypothesis has been formulated:

> It is expected that, accounting for the time and scope constraints of the project, a tenuous effect will be observed with respect to weather altering transportation patterns of users of the Dutch mobility system. Specifically, it is expected that as weather conditions get more extreme – that is, with respect to, rain and temperature – the percentage of bicycles used decreases.

Geographical and Temporal Scale

Having identified the research gap and formulated a hypothesis of the expected results, it is important to go into further detail about the intended geographical and temporal scale of the project.

On one hand regarding the geographical scale, this has been set at the national level for the Netherlands (overseas territories notwithstanding). This was done with the aim of observing transport patterns across the same transport system with as much weather variety as possible. More concretely, the geographical scale has been set to only account for urban areas as this is the only part of the Netherlands where easy access to data is possible.

On the other hand, the temporal scale was determined by ideating the amount of data that could be realistically accounted for with the scope of the TIL6022 project. From this basis, it was determined that a year's worth of data would be the ideal time period. Specifically, this was done for the year 2023 to ensure that the data itself was reliable and was less likely to need subsequent revision. The appropriateness of the temporal scale used has been further evaluated in the discussion.

Data Cleaning and Pre-Processing

In order to use the databases for subsequent analysis, these had to be cleaned and pre-processed to remove outliers and ensure that the data itself was reliable. Through such a process, the final datasets that were used for analysis could be reached.

1. The OV-fiets datasets for 2023 were obtained from the Onofficieel Archief Reisinformatie Nederlands Openbaar Vervoer (rijdendetreinen.nl) open data portal. After downloading, the files were saved locally and preprocessed to retain only the relevant variables — namely, the timestamp of the log, the station name, and the number of bicycles available. All other columns were removed to ensure consistency and reduce the size of the dataset and make it easier to interpret. The cleaned data was then converted into a parquet file – a binary compressed structure – for more efficient storage and faster processing.

   Subsequently, daily totals of rented bicycles were calculated for each station based on the difference in the number available bikes. It was consequently assumed that all bikes that were no longer available were rented out, which has been further evaluated in the discussion. By doing this, an estimation of the sum of daily bike rentals was possible, which in turn allowed for the calculation of weekly averages for OV fiets rentals per station. Furthermore, only the negative difference of each entry is included in the final sum used since that would reflect only the

decrease of the supply, hence only the bikes that are taken out for usage. It was important to calculate weekly averages to account for the cyclical weekly nature of bike rentals.

2. The KNMI (Koninklijk Nederlands Meteorologisch Instituut) dataset used in this study covers the entire year of 2023. The dataset includes meteorological observations collected from multiple weather stations across the Netherlands. From the available parameters, only the necessary columns such as temperature, precipitation, and weather station identifiers were retained.

   The data itself contains daily datapoints per location, with the data being the maximum temperature of that day per location and the precipitation being the total rainfall for that day per location. During preprocessing, datapoints were be converted to a weekly value by taking the highest value of the maximum temperature and one entry for the sum of the rainfall per week per location. By doing so, the cyclical factor of the day of week is removed so it can be used for proper evaluation.

3. For the TransLink dataset, only the essential columns were retained to focus the analysis on public transport usage patterns. Specifically, the dataset was filtered to include check-in counts, the timestamp (time of check-in), and date and hour information.

Data Integration

In order to allow for successful data integration, several steps were taken to allow for analysis between the OV Fiets bike rental data and the KNMI weather data, and the Translink and KNMI weather data. The steps taken for each process have been detailed below:

OV-Fiets Bike Rental Data and KNMI Weather Station Data

Firstly, code was written that processes the OV-bike rental data and KNMI weather station data by linking together the respective locations with the smallest distance. Specifically, the code calculates the geodesic distance to all weather stations, identifies the closest one, and records this information along with the distance in a

new dataframe for subsequent use. From this, the parquet file is read containing bike rental information and is related to the data obtained from the nearest weather station location. Below in Figure X the geographical distribution of the OV Fiets stations used is shown next to the KNMI weather stations that are selected using the code.

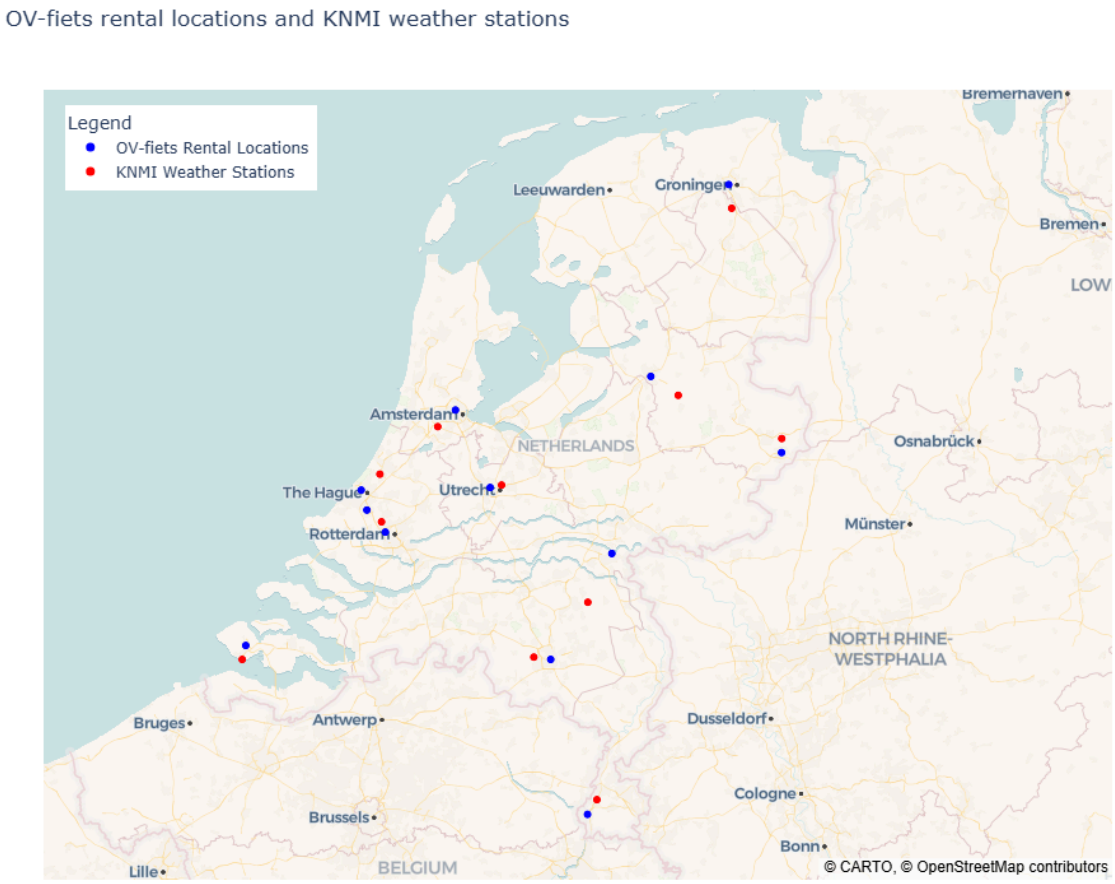

Figure 5 - Geographical locations of weather stations and OV-fiets locations that are used

Translink Data and KNMI Weather Station Data

Then, concerning the integration of data from Translink and KNMI, it is done by taking De Bilt as the reference value at a national level for temperature and precipitation. While it is correct to do so for temperature as it does indeed serve as a national reference value, the same cannot be said for precipitation, which is an inherently localised weather phenomenon. As such, this has been evaluated in the discussion.

Key Visualisations

In this section, linear regression is used to analyse the relationship between the bike usage and the two chosen weather conditions – maximum temperature and rainfall. The time components is removed by using the weekly data points described in Chapter "Data Cleaning and Preprocessing" so that the analysis can be performed. The results are shown in Plot X and Plot X. Furthermore, the linear correlation is represented by Pearsons correlation coefficient (r) and the coefficient of determination is shown by R^2. Explanation of the results can be found in the following chapter.

Public Transport Ridership vs. Weather (Translink x KNMI)

The plotted linear regressions are shown in Figure 6 for the maximum temperature and in Figure 7 for total precipitation.

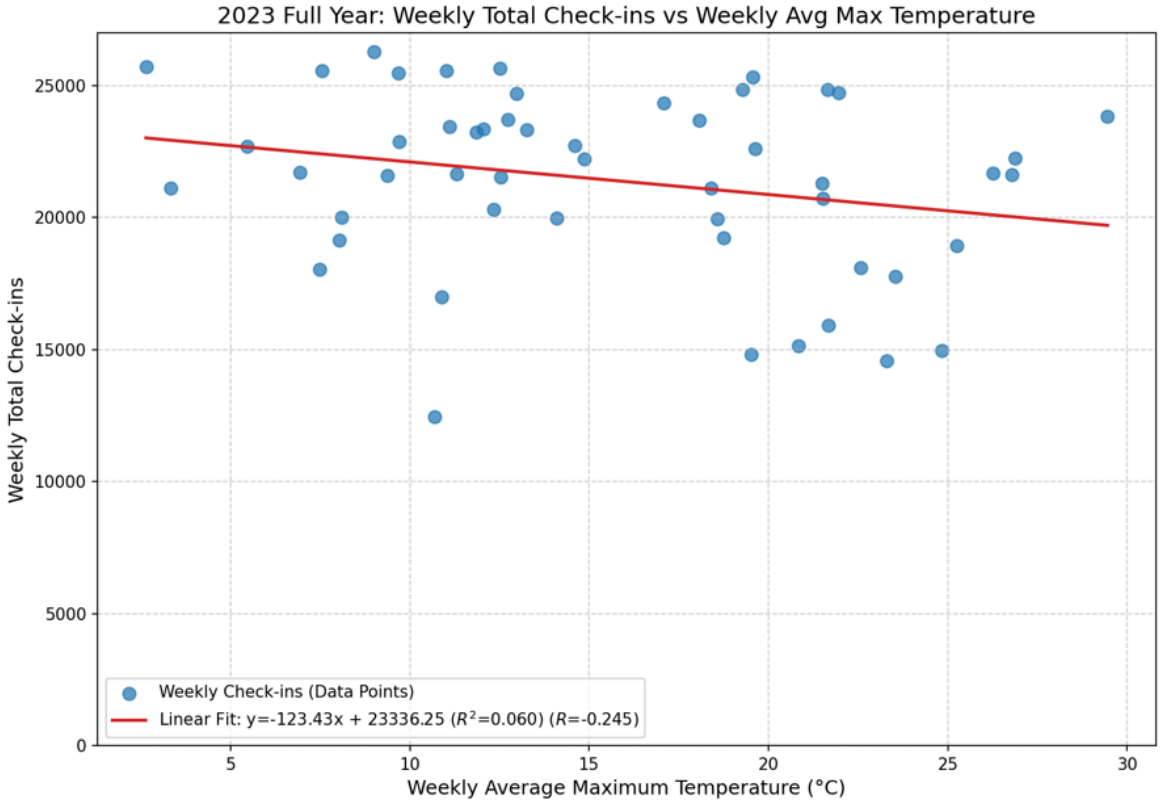

Figure 6 - Weekly total train check-ins and Weekly maximum temperature for 2023

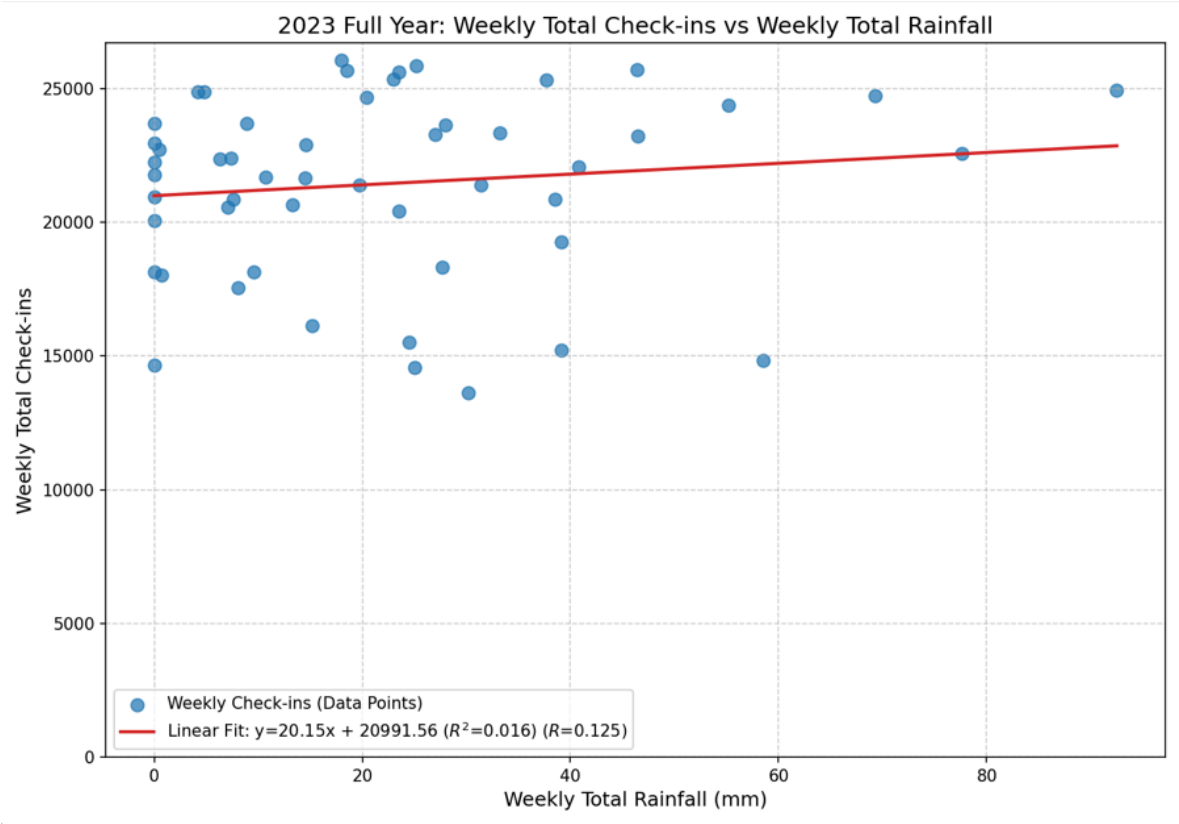

Figure 7 - Weekly total train check-ins and Weekly total precipitation

OV Fiets Usage vs. Weather (OpenOV x KNMI)

The linear correlation between OV-fiets dataset and the weather conditions are plotted on Figure 8 for maximum temperature and Figure 9 for total precipitation.

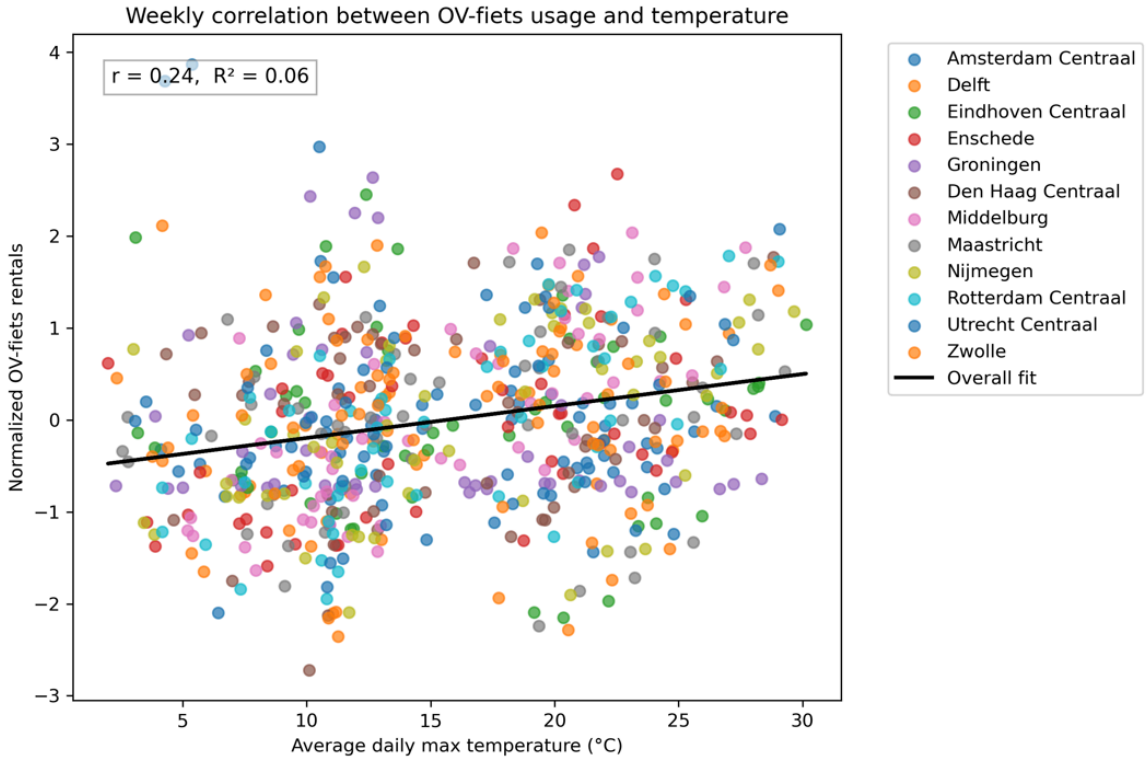

Figure 8 - Weekly correlation between OV-fiets usage and temperature

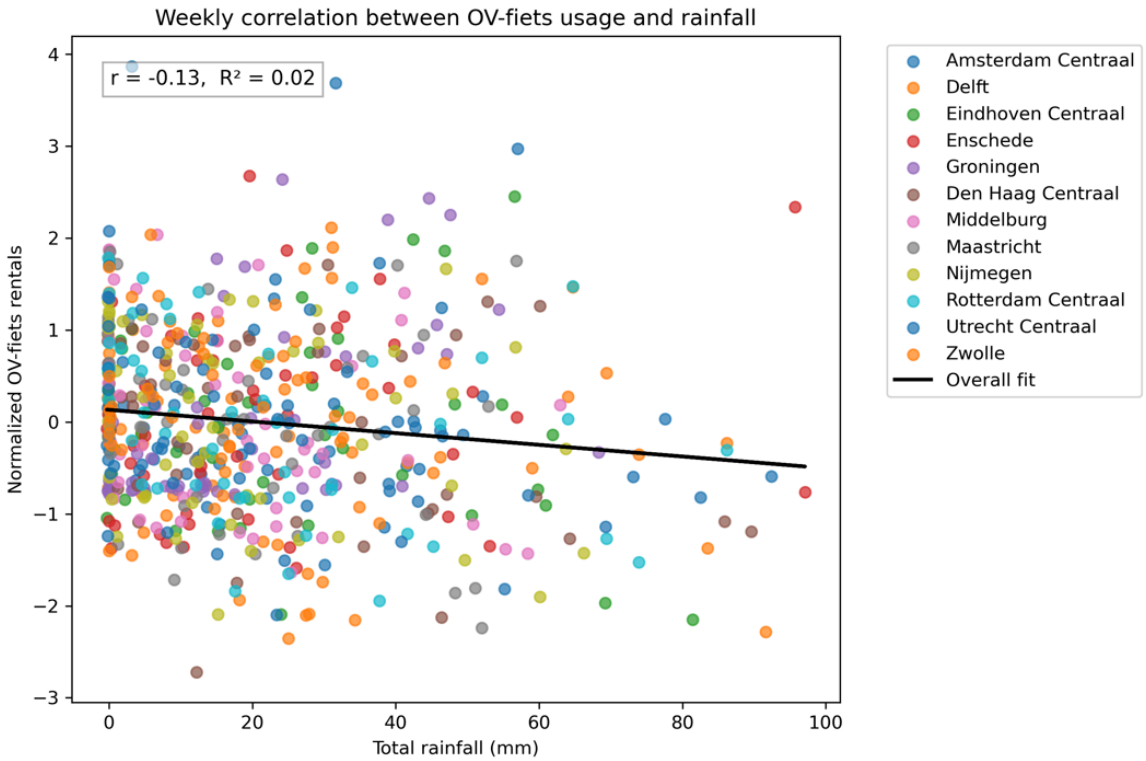

Figure 9 - Weekly correlation between OV-fiets usage and precipitation

Analysis of Results

Through the linear regression, several things can be evaluated, namely the linear correlation between the variables and the coefficient of determination of the data.

Linear correlation is represented by the Pearson correlation coefficient (r), which is a normalized covariance of the measurements from the data that fall in the range [-1; 1]. The measurements are normalized by calculating the ratio of the covariance of the two variables and the product of their standard deviations. Depending on the value of the Pearson correlation coefficient, the relation between the variables can be positively strong (close to 1), negatively strong (close to -1), or close to no relation (around 0).

On the other hand, R^2 represents the coefficient of determination of the data. It is used to evaluate the goodness of fit of the model of the statistical model. In other words, it shows how well the results represent the dataset as a whole. Since linear regression is used, R^2 will be calculated as the square of the Pearson correlation coefficient (r). It ranges from 0 to 1, with values around 0 point to a poor fit and values around 1 – to a good fit.

From plotting the linear regression of the public transport data and the weather conditions shown in Figure 6 and 7, the Pearson correlation coefficient show values of -0.245 for maximum temperature and 0.125 for total precipitation. Both have low influence on the weekly total check-ins with one being a positive relationship and the other – negative one. That is also showcased through the slope of the fitted line. As for the coefficient of determination R^2, the value for maximum temperature is 0.06, representing only 6% of the sample. Analogically, for total precipitation it is 0.016 or only 1.6%. This is the reason for the evident high dispersion of the data when compared to the fitted line.

The results of the other dataset shown in Figure 8 and Figure 9, on the other hand, showcase low Pearson correlations coefficients for both maximum temperature and total precipitation, 0.24 and 0.13 respectively. Therefore, both maximum temperature and total precipitation have relatively low influence on the OV-fiets bikes weekly usage. Through the linear correlation, a visible positive relationship is visible through the positive slope of the fitted line when comparing OV rentals to maximum temperature. However, when comparing total precipitation, the trend line has a negative slope, showcasing a negative relationship. Although both weather conditions have a visible relationship shown through the fitted line, the R2 values for both is low, pointing to a very poor fit. The maximum temperature has a R^2 of 0.06, representing only 6% of the total dataset, whereas R^2 of total precipitation is 0.02 or 2% of the total dataset.

Conclusion

Overall, it can be said that – based on the research conducted – a weak relation has been found between weather patterns and transport ridership patterns concerning train and OV fiets usage.

With respect to the research question posed, it can be said that a low correlation factors were found, with the highest being with respect to maximum temperature both for check-in's and OV fiets usage (with -0.245 between train check-ins and temperature, 0.125 between train check-ins and precipitation, 0.24 between OV Fiets usage and maximum temperature, and 0.13 between OV Fiets usage and precipitation). The comparison between the coefficients for Translink and OV-fiets ultimately was found to not make sense to do as the geographical scale differed for each analysis with Translink being compared on a national level and OV fiets on a city by city basis. Furthermore, a low R-squared was also observed for each of the correlation factors (0.06, 0.016, 0.06 and 0.02 respectively). Therefore, the amount of variance observed by the model is also quite low.

Looking back at the hypothesis formulated, it can be said that this hypothesis can be rejected through the data analysis conducted. Specifically, since it was hypothesised that a tenuous relation was to be observed, as per Magnusson (n.d.), a lower effect size was predicted. While it is true that this is the case for the correlation coefficients found, due to the low R-squared observed, it is not possible to state that these correlation coefficients are statistically significant. Furthermore, the research gap identified was not entirely possible to act on within the given scope of the project, where additional factors such as fog and wind were also identified to be analysed.

Following that, it can be concluded that more research is needed to evaluate whether the results obtained are statistically significant and capture the trends in the data.

Beyond this, however, it is important to mention that this project has helped refine the collaborative analytical skills of the team and served as a valuable experience for future data analysis projects required in the TIL masters.

Discussion

During the preprocessing and analysis parts, several discussion points arose. The discussion points are split into three parts, each one concerning a dataset used to conduct the analysis. Starting with the Translink database, the aggregation level of

the data is only on the national level. That leads to the inability to properly compare meteorological data with the train ridership. Currently, a compromise is made by using the weather station that is usually taken as the representation of the weather on a national level. That is, however, inaccurate and the analysis should be used with caution. If the aggregation level included check-in data per station, more precise and accurate analysis can be conducted. Furthermore, currently this dataset is used with the assumption that each check-in entry is corresponding to 1 traveller, which could not necessarily be the reality.

Moving on to the KNMI database, the data should be used with respect to location since some measurements used exhibit location-specific weather patterns, such as precipitation levels. Due to that reason, it is crucial to use weather data from the closest weather stations when comparing to the OV-fiets locations. That is sometimes not the case due to the low number of weather stations and their distance to certain locations. The code uses the closest possible weather station, however, it cannot correct the error that is created by the existing distance between the weather station and the OV-fiets location. Therefore, additional steps are needed to account for that existing distance, which is currently deemed out of the scope of this project and could be improved upon. Additionally, only temperature and precipitation is currently used, which is quite restrictive. Several different measurements can be included to expand the analysis, which were also identified in this project but they were later deemed out of scope.

Lastly, the OV-fiets database has several points that need to be discussed. From obtaining the data, it is stated that the data is obtained from the NS API. However, it is difficult to verify the data beyond taking it at face value and comparing it to other initiatives that use the data (e.g. https://ovfietsbeschikbaar.nl/). That poses a problem with reliability of the database as a whole since the information passes through a third party. Following that, the dataset is used with the assumption that each time the availability of bikes per location decreases, it represents a bike being rented out. In reality, a bike can be taken out of service for maintenance or other reasons, therefore, a precision of this measurement is somewhat questionable. Additionally, some bikes can be rented out for several days, which is currently impossible to reflect since there is no unique bike ID used in the database. The returning policy also allows for a bike to be returned to a different OV-fiets location. Although that case is uncommon, the possibility of that happening also introduces inaccuracies and uncertainties in the results. Concerning the structure of the database, it is difficult to optimise since a lot of data is provided all at once, with each entry being a periodical update of the amount of bikes at a specific station that is given every 15 minutes (NS, n.d.). That also plays a role in choosing the number of OV-fiets locations and the time period since it is required to load and to

go through the entire database, filter out unused locations and convert it to the necessary file format for more efficient loading. This whole process takes a lot of time and computing power.

Concerning the interaction between the datasets, due to how KNMI and OV-fiets datasets are related, it is impossible to do an analysis in rural settings due to the low number of weather stations and their distance to the OV-fiets locations in rural areas. As for the Translink dataset, it is already highlighted that the difference of aggregation level poses problems when compared to the KNMI dataset.

References

Böcker, L., Dijst, M., & Faber, J. (2016). Weather, transport mode choices and emotional travel experiences. Transportation Research Part A: Policy and Practice, 94, 360–373. https://doi.org/10.1016/j.tra.2016.09.021

Böcker, L., Priya Uteng, T., Liu, C., & Dijst, M. (2019). Weather and daily mobility in international perspective: A cross-comparison of Dutch, Norwegian and Swedish city regions. Transportation Research Part D: Transport and Environment, 77, 491–505. https://doi.org/10.1016/j.trd.2019.07.012

de Kruijf, J., van der Waerden, P., Feng, T., Böcker, L., van Lierop, D., Ettema, D., & Dijst, M. (2021). Integrated weather effects on e-cycling in daily commuting: A longitudinal evaluation of weather effects on e-cycling in the Netherlands. Transportation Research Part A: Policy and Practice, 148, 305–315. https://doi.org/10.1016/j.tra.2021.04.003

Galich, A., & Nieland, S. (2023). The Impact of Weather Conditions on Mode Choice in Different Spatial Areas. Future Transportation, 3(3), 1007–1028. https://doi.org/10.3390/futuretransp3030056

Magnusson, K. (n.d.). Understanding Statistical Power and Significance Testing — an Interactive Visualization. Rpsychologist. Retrieved November 7, 2025, from https://rpsychologist.com/d3/nhst/

Ministry of Infrastructure and Water Management. (n.d.-a). KNMI - Daggegevens van het weer in Nederland. Www.knmi.nl. Retrieved October 3, 2025, from https://www.knmi.nl/nederland-nu/klimatologie/daggegevens

Ministry of Infrastructure and Water Management. (n.d.-b). The Netherlands Mobility Panel. Mpndata.nl. Retrieved October 3, 2025, from https://mpndata.nl/

Sabir, M., van Ommeren, J. N., Koetse, M. J., & Rietveld, P. (2010). Weather and travel time of public transport trips: an empirical study for the Netherlands. In M. Givoni & D. Bannister (Eds.), Vrije Universiteit Amsterdam (pp. 275–288). Routledge. https://hdl.handle.net/1871.1/46e29e14-b0d2-480a-8eeb-bf9f94004e3c

Translink. (n.d.). OV-Data: Open Data. Translink.nl. Retrieved October 3, 2025, from https://translink.nl/open-data/

van Natijne, A. (2015). Onofficieel archief reisinformatie Nederlands Openbaar Vervoer. Fwrite.org. https://trein.fwrite.org/idx/dedup_OVFiets.html

Wilkesmann, F., Ton, D., Schakenbos, R., & Cats, O. (2023). Determinants of station-based round-trip bikesharing demand. Journal of Public Transportation, 25, 100048–100048. https://doi.org/10.1016/j.jpubtr.2023.100048