[redacted]

Data Science Project

Badi-data

Conceptual Design Report

31 October 2023

Abstract

During some hot summer days PubliBike struggles with the logistics of providing each station with enough bicycles, as the hot weather invites many people looking to cool down for a so called “Aare-Schwumm”, a swim in the local river, Aare.

By using the hourly entrances into the outdoor swimming pool Marzili, the hourly air temperature, the hourly water temperature of the Aare and the hourly PubliBike availabilities of selected stations during the summer season 2023 we are trying to predict the demand for PubliBikes and the expected number of visitors during a day by inputting the forecasted hourly air and water temperatures, the week day and special events.

Table of Contents

1. Project Objectives 2
2. Methods 3
    2.1 Infrastructure 3
    2.2 Software Libraries and Tools 3
    2.3 Modeling, Algorithms & Statistical Methods 3
3. Data 4
    3.1 Data Collection and Acquisition 4
        Air temperature 4
        Water temperature 5
        Number of entrances into the outdoor swimming pool Marzili 5
        PubliBike availability 5
    3.2 Data Cleaning and Preprocessing 5
    3.3 Data Collections - Overview 6
        Air temperature 6
        Water temperatures 6
        Number of entrances into the outdoor swimming pool Marzili 7
        PubliBike availability 7
4. Metadata 7
5. Data Quality 7
6. Data Flow 8
7. Data Model 8
    7.1 Conceptual 8
    7.2 Logical 9
    7.3 Physical 9
8. Documentation 10
9. Risks 10
10. Conclusions 10

1. Project Objectives

During some hot summer days PubliBike struggles with the logistics of providing each station with enough bicycles, as the hot weather invites many people looking to cool down for a so called “Aare-Schwumm”, a swim in the local river, Aare. According to a news article (“Hitzetage in Bern – Hohe Temperaturen sorgen für Dichtestress an der Aare”)[1] many people drive their PubliBike to the Eichholz station, leave it there and swim down to the Marzili. However, even during the "non-swimming season", it appears that certain stations experience a shortage of available bicycles at specific times.

We want to check whether there is a connection between the air temperature in Bern, the water temperature of the Aare, the daytime, the weekday, and special events (i.e. bank holidays or school holidays) on the amount of PubliBikes that are available in the city of Bern and further on the number of entrances in the outdoor swimming pool Marzili.

The project therefore has a dual objective: firstly, to predict the probability of PubliBike availability at a specific station at a given time, and secondly, to anticipate the number of visitors at Marzili. This twofold approach is based on the assumption of a strong correlation between bike availability and visitor numbers.

Therefore, we can formulate the two following objectives:

Primary Objective:

Predict the probability of the bike availability at PubliBike stations, considering the hourly air and water temperature, weekday, and special events.

Secondary Objective:

Predict daily and hourly entrances at Marzili open swimming pool based on expected air and water temperatures for the next day, the weekday, and special events.

[1] “Hitzetage in Bern – Hohe Temperaturen sorgen für Dichtestress an der Aare.” Der Bund, 24 August 2023, https://www.derbund.ch/hitzetage-in-bern-hohe-temperaturen-sorgen-fuer-dichtestress-an-der-aare-245412356738

2. Methods

2.1 Infrastructure

The data will be stored on Google Drive after the initial collection from the different sources. By mounting the Google Drive to the Google Collab, we can access the data through our Python script which is stored on the Google Drive as well. The future data collection can be done by web-scraping the different pages.

2.2 Software Libraries and Tools

The chosen programming language for this project is Python. We expect to use many different libraries which are listed and shortly explained hereafter:

- Pandas and NumPy: for data manipulation such as data cleaning, preparation and exploration, and the descriptive statistics. The Pandas library will also be of big support with the time series data. NumPy on the other hand can handle data with more than two dimensions.

- Matplotlib and Plotly: for data visualization already in the early stages of the data exploration process. Plotly allows more interactive visualizations.

- scikit-learn & TensorFlow: the machine learning libraries will help us find the relations between dependent and independent variables and will support us with some of the deep learning tasks.

- SciPy & Statsmodels: for time series modeling and hypothesis testing.

- GeoPandas: for geospatial data.

- BeautifulSoup & Scrapy: for web scraping and APIs

2.3 Modeling, Algorithms & Statistical Methods

For our predictions we will start with the linear regression models before going to more sophisticated time series models such as the AutoRegressive Integrated Moving Average (ARIMA). Further, we will also apply Random Forest and Neural Network methods to our data, to see which method gives us the best results.

We will need to do the following data feature engineering:

- The day of the week: this feature might have an impact on the prediction.
- Special days or events: during bank holidays or the school children's holidays we can expect a different outcome compared to the regular working day outside of the holiday season.

The Accuracy or the Precision and Recall is an appropriate evaluation metric for the primary objective and the Mean Absolute Error, Root Mean Squared Error or the R-squared are good measures to evaluate how good the predictions are.

An 80:20 time-based split seems appropriate to validate the data.

3. Data

3.1 Data Collection and Acquisition

To train our model we want to use the following data sets, covering a specified time range (summer season 2023). Data collection should occur at an hourly frequency and encompass the following parameters.

- Air temperature
- Water temperature
- Number of entrances into the outdoor swimming pool Marzili
- PubliBike availability

Air temperature

The hourly air temperatures will be provided by MeteoSwiss. MeteoSwiss operates stations at different locations. We consider the location, Bern, Bollwerk as the most appropriate location.

The data can be downloaded directly through the data portal for teaching and research (IDAweb) which is maintained by MeteoSwiss[2].

[2] MeteoSwiss. 9 March 2019, https://gate.meteoswiss.ch/idaweb/login.do.

Water temperature

For water temperature information, we will rely on the Federal Office for the Environment (FOEN), with measurements taken at the Schönau location[3]. Since historical water temperature data over an extended period isn't publicly accessible, it will need to be requested from FOEN.

Number of entrances into the outdoor swimming pool Marzili

The count of entrances at the Marzili outdoor swimming pool is recorded by the Sportamt der Stadt Bern, and this data needs to be requested from their department.

PubliBike availability

PubliBike availability at different stations is monitored by PubliBike itself. Access to historical PubliBike availability data is restricted to the public, necessitating a direct request to PubliBike. While current availability can be scraped online[4], the initial dataset acquisition will be through the company. For future updates, scraping directly from online sources can be considered, given that current PubliBike availabilities, air temperatures, and water temperatures are publicly accessible online

3.2 Data Cleaning and Preprocessing

The datasets may contain missing values, which would lead to a lower performance of our model. To ensure that the data quality is high, we will identify the missing values at the beginning by doing descriptive statistics and visualizations. Where appropriate, we will induce proper imputation techniques by using logical thinking and our domain knowledge to ensure that we are working with the largest possible data set to train the model.

For instance, when dealing with missing data related to variables such as air and water temperature, we initially assume complete data. However, if any gaps emerge, we explore strategies such as leveraging measurements from the nearest available location to fill these voids.

On the other hand, when addressing missing data concerning the availability of PubliBikes, our domain knowledge may not provide a basis for informed imputation. In such cases, we resort to conventional imputation methods.

---

[3] Federal Office for the Environment (FOEN). “Aare - Bern, Schönau (2135).” *Hydrodaten.admin.ch*, https://www.hydrodaten.admin.ch/de/seen-und-fluesse/stationen-und-daten/2135

[4] PubliBike. “publi - e bike availability bern.” *opendata.swiss*, 3 March 2023, https://opendata.swiss/en/dataset/publi-e-bike-availability-bern.

3.3 Data Collections - Overview

This chapter gives a very first overview of the already gathered Data.

Air temperature

Figure 1 displays the hourly air temperature measurements taken at Bern, Bollwerk. This visualization illustrates that the dataset exhibits neither conspicuous errors nor a substantial quantity of missing values.

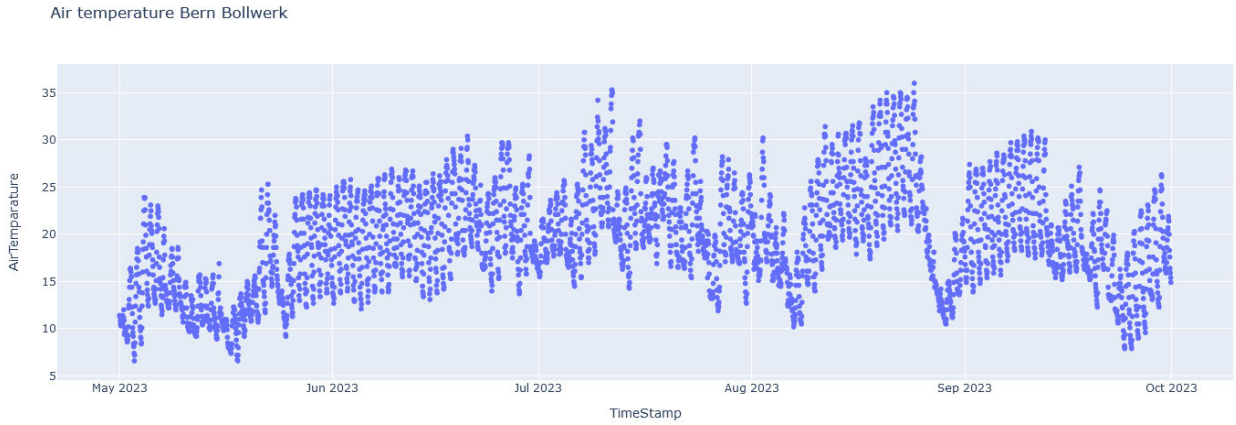

Figure 1: Time series of hourly air temperature at Bern, Bollwerk

Water temperatures

Figure 2 depicts hourly water temperature measurements recorded in Schönau. This visualization confirms the dataset's accuracy, showing no significant errors or notable gaps in the data, like the air temperature data.

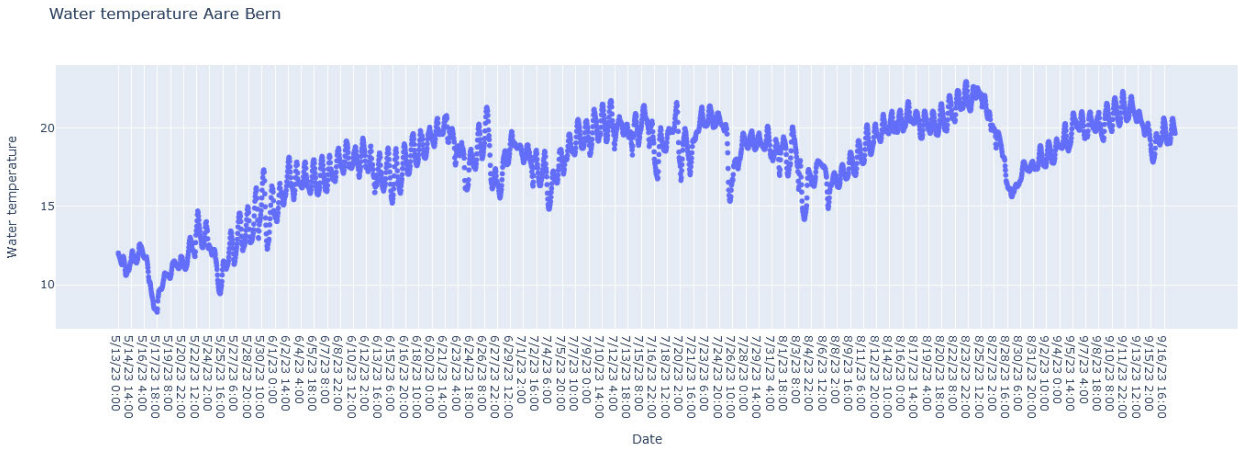

Figure 2: Time series of hourly water temperature at Schönau

Number of entrances into the outdoor swimming pool Marzili

Figure 3 displays the hourly visitor count at Marzili. Once again, there are no apparent issues with the data.

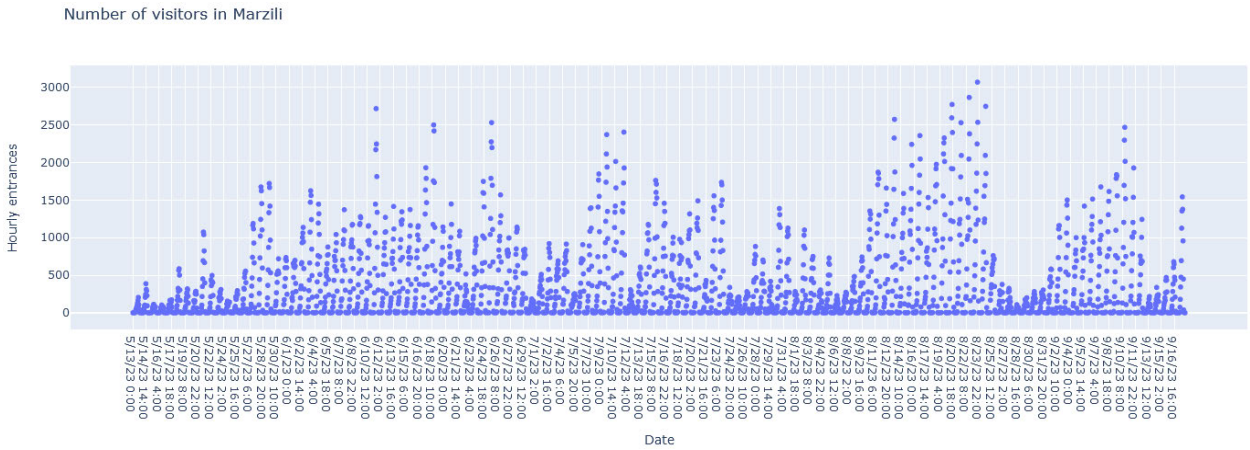

Figure 3: Time series of hourly visitor counts at Marzili.

PubliBike availability

Getting the Publibike availability data is still ongoing. But research that has been done so far shows that there will be Data that shows the availability measured every 10 minutes. And both the quality and quantity of the data seem promising. Therefore, if Publibike provides us with the data, we anticipate no major issues.

4. Metadata

We are utilizing the data described in Chapter 3, which includes air and water temperature, the number of visitors at Marzili, and the availability of PubliBikes at a station. This data covers the specified time period from May 15, 2023, to September 15, 2023, with measurements taken every hour. If a third party wishes to access this data, they have two options: they can request it directly from the data source, as we did (see Chapter 3.1), or they can download it from our Google Drive.

5. Data Quality

After the cleaning and preprocessing steps outlined in Chapter 3.2, we anticipate good data quality. However, two issues may significantly impact our dataset.

The first issue pertains to the number of entrances and the availability of PubliBikes. We only have data for the summer season of 2023. This limitation arises because the Sportamt der Stadt Bern initiated the counting of entries in 2023, and PubliBike can provide data starting from May

2023. We have concerns that this relatively short time period may not be sufficient for a comprehensive analysis.

The second, though less critical issue, relates to certain days when the entrance counter at Marzili experienced problems. Consequently, for these specific days, we lack any data at all.

6. Data Flow

The following data flow model shows the different steps of the project:

7. Data Model

7.1 Conceptual

At the conceptual level we outline the high-level idea of the project.

By taking the expected air and water temperatures in Bern on a given day we can predict the PubliBike availability at each station and in a next step predict the expected visitors into the Marzili swimming pool each hour in a day.

We can then create an app that helps people plan their journey by showing them the probability of finding a PubliBike at a specific station during a specific hour. Further we can sell our model to PubliBike to help them with planning the logistics of their bicycles on a given day, given the expected temperatures.

For the Marzili operators this model could help them optimize the personnel planning.

7.2 Logical

At the logical level, we define the specific columns/features that we will use for our analysis and the modeling.

For the temperature data we will need the following features:

- Timestamp (DateTime)
- Air Temperature (Float)
- Water Temperature (Float)

For the PubliBike availability data we will need those features:

- Timestamp (DateTime)
- Station ID (Integer)
- Available Bikes Count (Integer)

For the entrances data we will use the following features:

- Timestamp (DateTime)
- Entrances Count (Integer)

From the Timestamp feature the following additional features will be created:

- Day of Week (Integer)
- Month (Integer)
- Calendar Week (Integer)
- Special Events (Binary Indicator)

7.3 Physical

At the physical level, we consider the infrastructure and tools needed to store and process the data. As we are not working with very big data sets, we are expecting that we can store the data and the model on google drive and therefore we have no greater requirements in this perspective.

8. Documentation

As we are using Google collab, we will be commenting on the script, so that the code can be replicated for other similar projects.

9. Risks

The main risk that we are seeing is in the data collection process of the PubliBike data. This data is not publicly available and is owned by a private company. All the other data is either owned by the federal government or the city and should therefore be accessible for everyone. If we do not get the PubliBike information, then we would need to terminate our project for this year, but we can collect data in the future by scraping the current PubliBike availability data and by storing it on our servers. This means that we can only do our project a year later.

10. Conclusions

In this conceptual design report we outlined that our two objectives of the project are to predict the hourly PubliBike availability at specific stations and the hourly and daily entrances into Marzili based on the expected hourly air and water temperatures, the week day and special events like bank holidays or school holidays. Further we showed how we are intending to achieve those objectives. The data that we collected so far are of good quality and we expect to have good results with the models we intend to use.

We are confident that our two project objectives can be achieved, we just don't know yet whether we will be able to build and train our model with the PubliBike availability data for the summer season 2023, as we did not receive any feedback from PubliBike yet. As stated earlier in the report, there is the possibility to scrape current PubliBike availability data from the web, so for the future we could collect the data in this way. Regardless of whether we receive the data from PubliBike, we expect that we will be able to start the project and hopefully achieve the secondary objective for the summer season 2023 already and then in the worst case use the scraped data in 2024 to achieve the primary objective next year.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 31.10.2023

References and Bibliography

[1] Federal Office for the Environment (FOEN). “Aare - Bern, Schönau (2135).” *Hydrodaten.admin.ch*, https://www.hydrodaten.admin.ch/de/seen-und-fluesse/stationen-und-daten/2135.

[2] “Hitzetage in Bern – Hohe Temperaturen sorgen für Dichtestress an der Aare.” *Der Bund*, 24 August 2023, https://www.derbund.ch/hitzetage-in-bern-hohe-temperaturen-sorgen-fuer-dichtestress-an-der-aare-245412356738.

[3] MeteoSwiss. 9 March 2019, https://gate.meteoswiss.ch/idaweb/login.do.

[4] PubliBike. “publi - e bike availability bern.” *opendata.swiss*, 3 March 2023, https://opendata.swiss/en/dataset/publi-e-bike-availability-bern.

[5] Sportamt der Stadt Bern. “Badi-Bilanz 2023: Wechselhafte Saison mit Rekordtagen - Sportamt der Stadt Bern – Bern bewegt!” *Sportamt Bern*, 8 September 2023, https://www.sportamt-bern.ch/badi-bilanz-2023-wechselhafte-saison-mit-rekordtagen/.