



---


**Predicting the number of cyclists in streets of the city  
of St. Gallen**  
Research Report

---

Group 05

Nina 

Luc 

Tonio Isenschmid 

## 1 Introduction

Cycling has the potential to become *the* primary mode of short distance transportation – also in hilly towns like St. Gallen. Well-developed cycling infrastructure, including wide lanes and bike-friendly urban planning, encourages and facilitates cycling despite the terrain. Moreover, cultural shifts towards sustainable and healthier lifestyles contribute to the prominence of cycling. Finding out, where the most frequented cycling areas lie could be an important factor in facing the challenges emerging due to this change.

The aim of this project was to use cycling and weather data, which has been collected over the past decade, to predict the number of cyclists in the streets of St. Gallen by applying several regression models.

## 2 Data collection and preprocessing

The findings of this research report are rooted in the bike and weather data, forming the fundamental basis of the analysis. This section will focus on the process in which the data for this project was gathered and how it was altered to be useful the overall results.

### 2.1 Data sources

The report made use of four different data sources including official state-run webpages and multiple event publications. Therefore, the data sources enabled us to realise a project in a local context.

Firstly, both the weather and the cycling dataset were obtained from the official data publication site of the canton of St. Gallen. Whilst the cycling dataset was created and published by the *Tiefbauamt* of the city of St. Gallen, the weather dataset was generated and issued by the *Swiss National Basic Climatological Network*. The cycling dataset contains approximately 39'000 observations in up to 15 different locations between 2011 and 2021. Its features include date, weekday, and the exact location of the observations, while the label represents the measured number of cyclists on the given day. The weather dataset stands for weather statistics such as temperature, sunshine duration, humidity, snowfall, and rain, recorded for dates spanning from 2011 to 2022. Finally, the project incorporates self-collected data on St. Gallen's top five annual events and the dates of official state holidays. This includes the exact days on which the Open Air St. Gallen, OLMA, St. Gallen Fest, Symposium or CSIO horse riding competition have taken place as well as the exact days of official holidays.

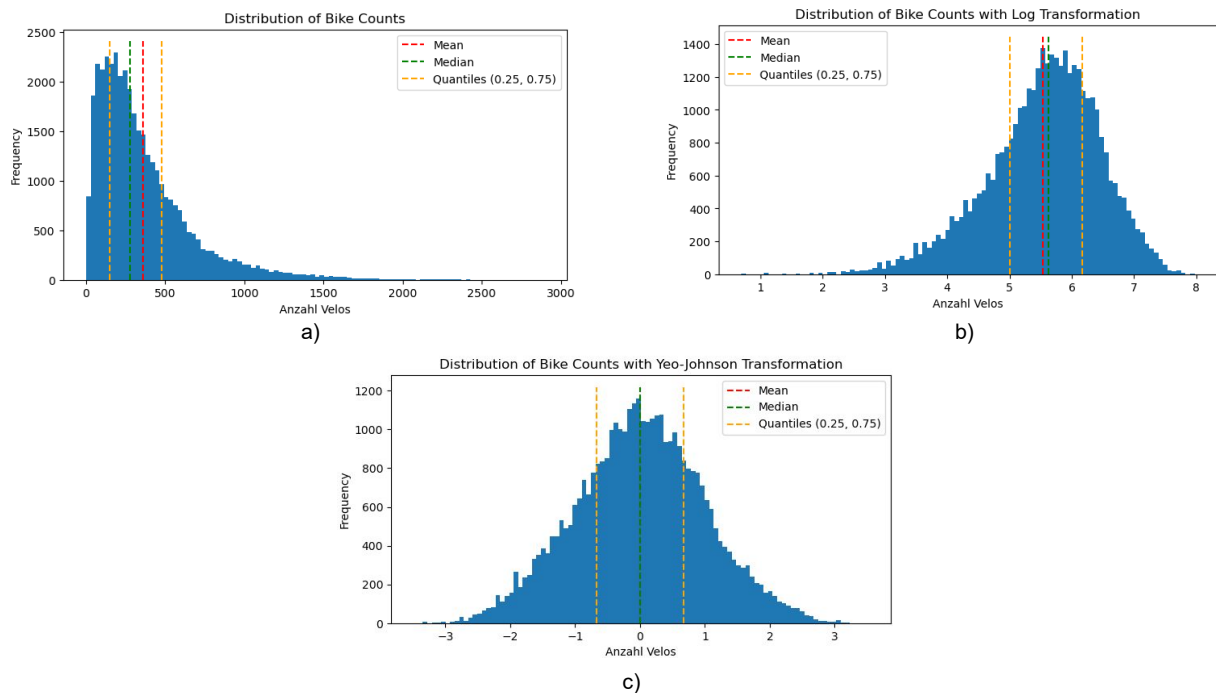
### 2.2 Data Visualisation and Preprocessing

The datasets were first rearranged to represent an ascending timeline. In addition, the columns (location coordinates, day number, key figure and working day) were dropped from the cycling dataset, because they would have not been of great use to the overall goal of our prediction

model. Meanwhile, the weather dataset was sorted and refined to align with the dates in the cycling data. After having sorted the values, the datasets were merged by trying to retain as many rows as possible. To implement the exact weekday, we used *dummies* to add seven indicator columns to the dataset, instead of just one column containing seven different entries. The now combined dataset was supplemented with additional indicator columns of the five previously mentioned events and school holidays from the year 2011 to 2021. Furthermore, we addressed missing values, in our case represented by 0s. Since they constituted a statically irrelevant amount with approximately 1.4% of total observations, we decided to simply drop them.

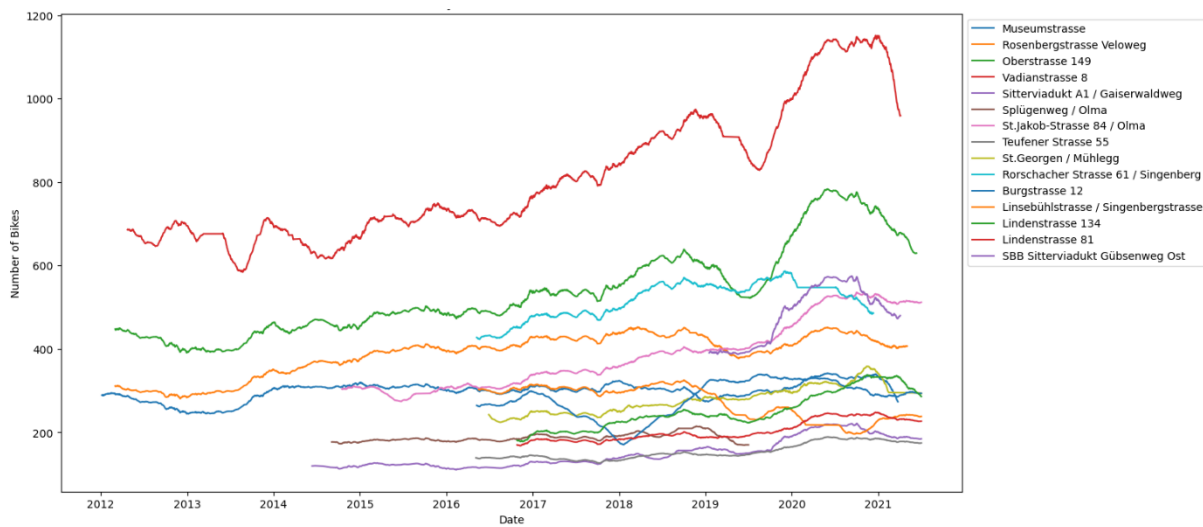
Prior to applying models, we assessed the label's underlying distribution, considering that parametric methods tend to handle nonlinear relationships less effectively. The following figure 1a displays the distribution of the number of bikes, which is clearly right-skewed. As this could potentially influence models, *Logarithmic* (b) and *Yeo-Johnson Transformations* (c) were applied to approximate a normal distribution. The latter out of the *Power Transform Family* seems to be working well for now because it balances variance and reduces skewness.

**Figure 1: Distribution of Total Bike Counts**



Our dataset contains 15 different observation locations all around the city of St. Gallen. Not only do the locations inhibit diverse durations and trends as shown on figure 2, we also strongly suspect there to be distinct correlations between the features and the outcome variable *Anzahl Velos* for each location. Hence, we trained a model for every location by creating sub-data frames.

**Figure 2: Yearly Trends for All Locations**



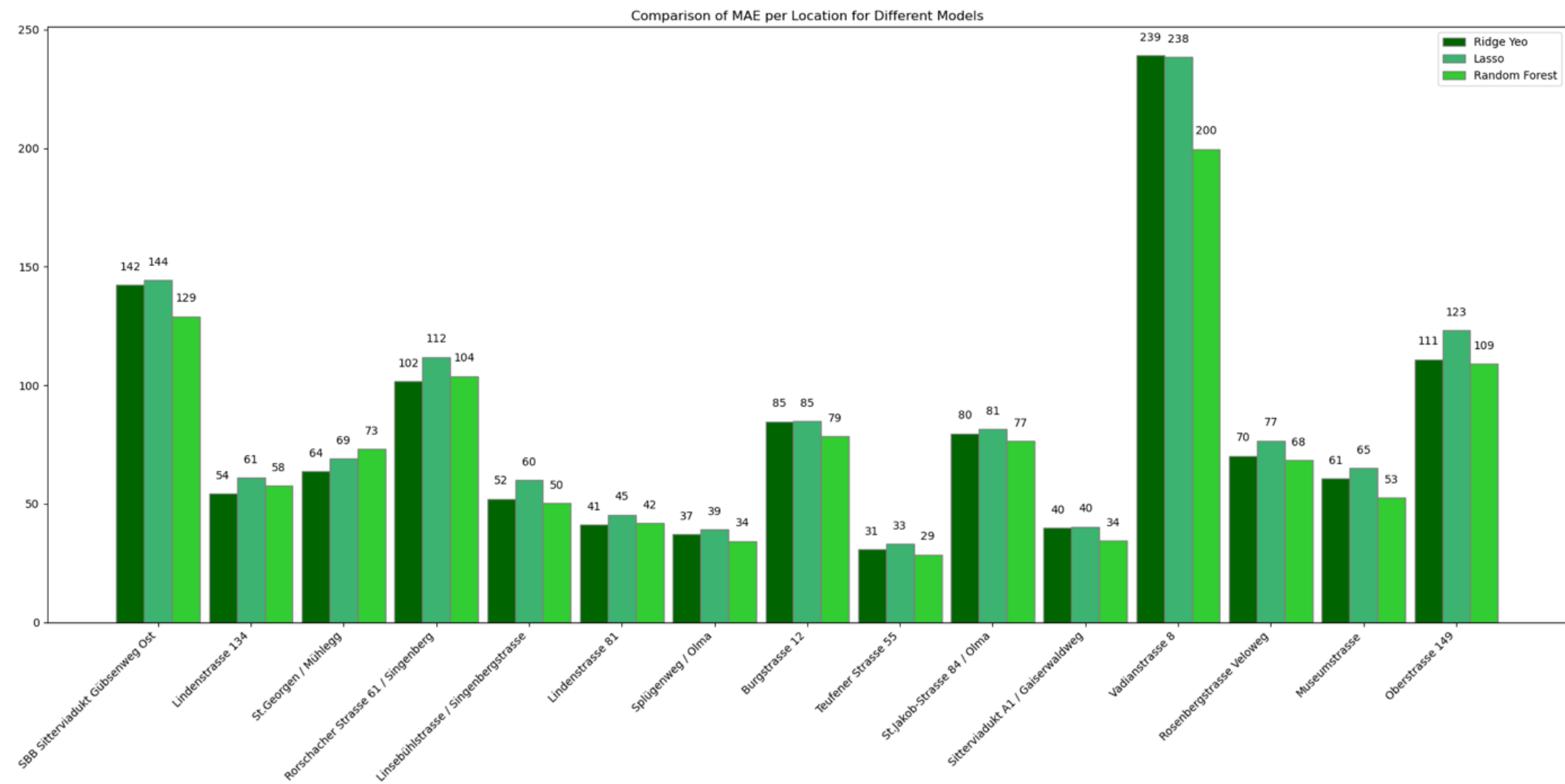
Firstly, the possibly varying correlations might be stemming from differences in geographics. One street may be flat, another one busy with a cycling lane, and a third situated on a downhill slope from Rosenberg. Naturally, rain or snow would affect the number of cyclists differently at these diverse locations. Secondly, having different models for various locations allows for better interpretation. For instance, we had only realised that the 0s are a problem after having created a model for every location and wondering why some locations exhibited terrible MAEs and MSEs. Lastly, we can avoid trouble with *Time Series Split* due to multiple entries for one date if we train models for every location.

### 3 Model Selection

With the *bias-variance tradeoff* in head, we acknowledged *interpretability* by utilising linear models such as *Ridge and Lasso Regression*. Simultaneously, we explored *Random Forest Regressor* as a non-parametric method, focusing on its *flexibility* to capture complex patterns without relying on specific assumptions about the data's underlying distribution. To evaluate our models' performance, we used *Times Series Split* as our cross-validation method, similar to the k-fold method but considering the temporal aspect. *Ridge and Lasso Regression* were applied across various transformations – none, *Log*, and *Yeo-Johnson*. Meanwhile, the *Random Forest Regressor* was employed without any transformation due to its adeptness in handling non-linear relationships.

For each model, we opted for the most suitable transformation method based on achieving the lowest average MAE and MSE. Specifically, we utilized Yeo-Johnson transformation for *Ridge* and omitted transformation for *Lasso*. After evaluating the three models, we chose to proceed with the *Random Forest Regressor* due to its ability to yield the lowest MSE and MAE.

**Figure 3:** Comparison of the MAEs for Ridge Yeo-Johnson, Lasso and Random Forest



## 4 Interpretation

As visible in figure 2, there has been a drop in the trend of number of bikes for certain locations during 2019, such as *Vadianstrasse*, *Oberstrasse* and *Teufener Strasse*. All those streets lead into the area of *Roter Platz*, where construction works were taking place during 2019 ([Tagblatt](#)). The same accounts for *Rosenbergstrasse*.

Moreover, explaining the subtle performance discrepancies between Lasso and Ridge Regression models proved to be a challenging task. The reason must lie within the way Ridge and Lasso treat their coefficients, i.e., shrinking them towards, respectively exactly equal to zero. However, the exact reasoning would be beyond our capabilities. Without a doubt, it can be stated that Random Forest Regression significantly outperforms the linear models. Whilst Ridge considers all features and weights them differently, the tree-based model neglects a great part of predictors altogether. This is to ensure that strong predictors do not overly dominate, thereby preventing the bagged trees from looking too similar. Further, their random sample method prevents from overfitting.

## 5 Reflection

In retrospect, we recognised the significance of investing time in refining our dataset. A more focused approach towards data inspection could have simplified our analysis, reducing the overall time invested in the process.

A case in point would be, that the number of values equaling zero was not considered relevant in the early stages of the project. Later, when computing the MSEs and MAEs, the impact was shown by a high discrepancy between the different locations. At this point, the importance of managing the values equaling zero became obvious, especially since half of all values equaling zero were from a single location. Although, by managing these values later, the impact of these extreme values could have been already foreseen when cleaning and preprocessing the datasets. In addition, only after computing the MSEs and MAEs of the whole dataset did the question arise whether the project should focus on the different locations separately, or on the whole dataset.

While the objective of this project led prediction modelling to be taken in a more specific approach to each individual location, computing overall predictions could be a matter of further research. In such a particular instance, the need for more complete data that would be gathered over longer periods of time could have an incremental effect on predictions.