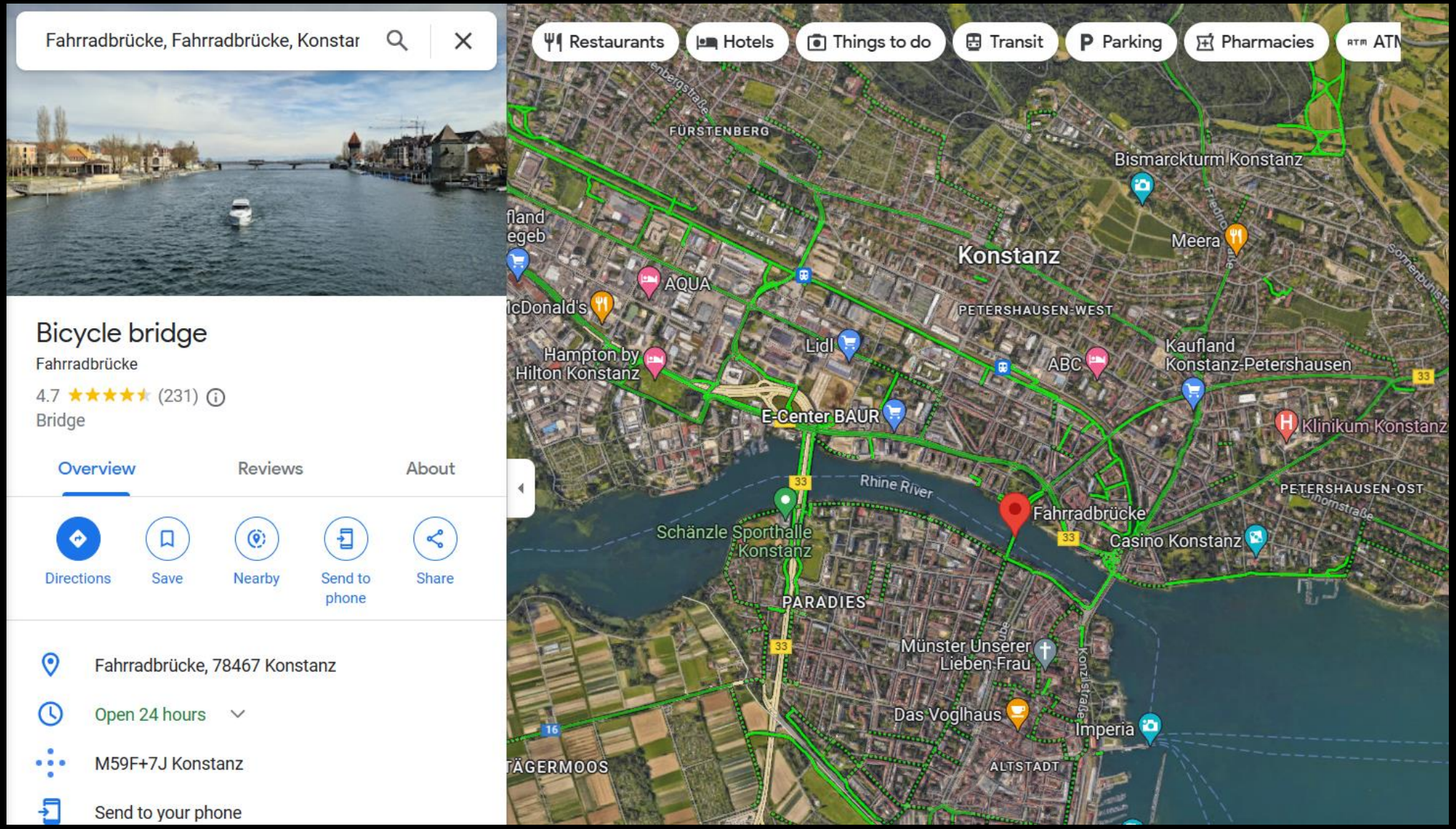
Data Analysis on the Total Bike Flow on the Konstanz Bicycle Bridge and Bike-Involved Accidents

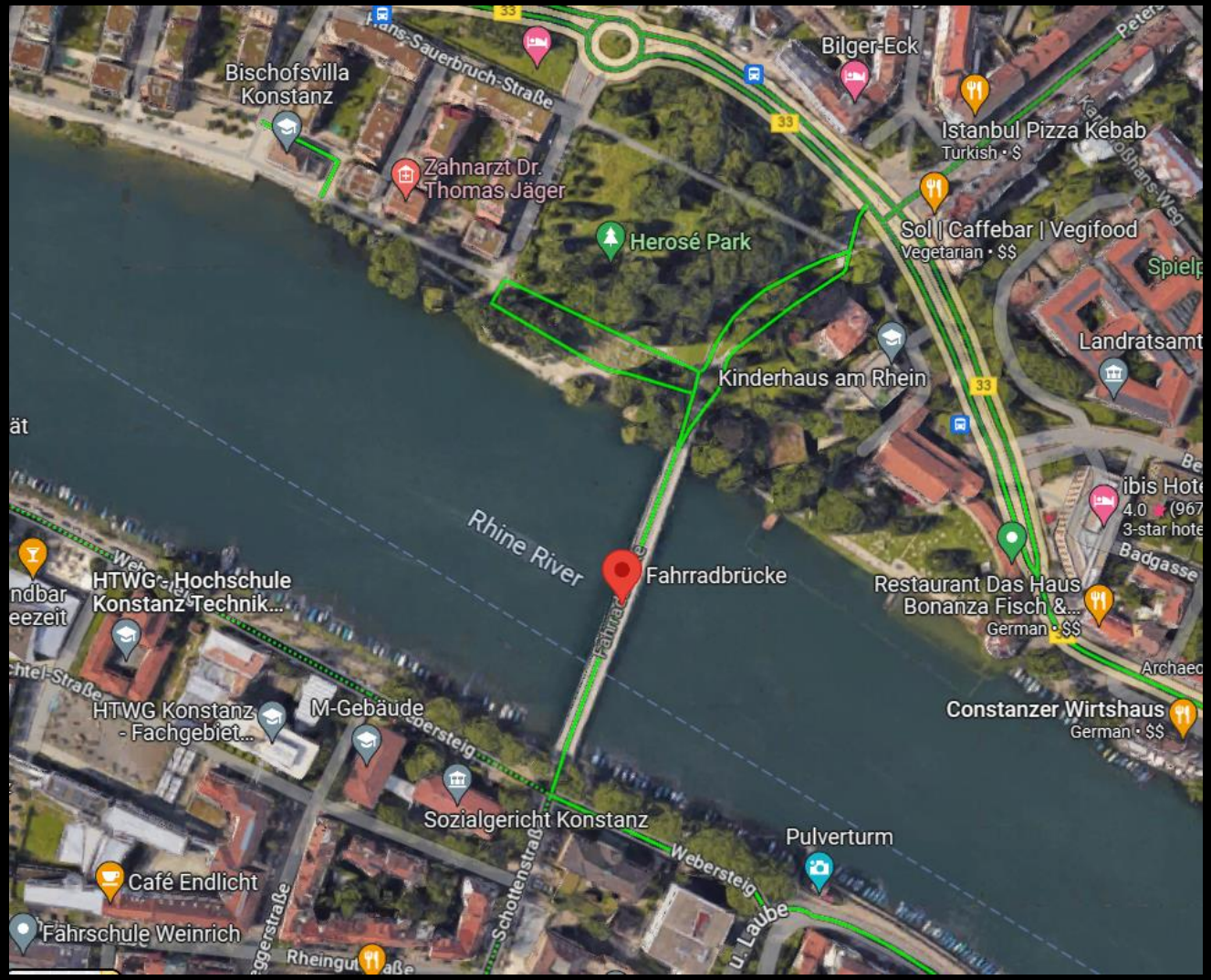
Osman Yigit

GitHub Page: notmuchnerdy/2023-amse-template

The City of Konstanz and Bicycle Bridge



Bicycle Bridge



Bicycle Bridge has a spectacular location at the heart of Konstanz and It binds two parts of the city.

Unfortunately, there was no data about the overall bike traffic in Konstanz. Therefore, I assumed that the total bike flow towards each side of the ‘’Fahrradbrücke’’ can be perfectly representative of the total bike flow in the city.

Overview on the Bicycle Bridge Dataset

	Zeit	Fahrradbruecke	Fahrradbruecke stadteinwaerts	Fahrradbruecke stadtauswaerts	Symbol Wetter	Temperatur (°C)	Gefuehlte Temperatur (°C)	Regen (mm)
0	2020-01-01 00:00:00	104.0	31.0	73.0	Leicht bewoelkt	1.0	-1.0	0.0
1	2020-01-01 01:00:00	128.0	47.0	81.0	Leicht bewoelkt	0.0	-1.0	0.0
2	2020-01-01 02:00:00	178.0	77.0	101.0	Sonnig	0.0	-2.0	0.0

8785 rows × 12 columns

The bicycle bridge dataset contains information about the bicycle flow of each side of the city (Fahrradbruecke stadteinwaerts - Fahrradbruecke stadtauswaerts) and total bicycle flow (Fahrradbruecke). In addition to these columns, weather-related data also is involved (Symbol Wetter, Temperatur, Gefuehlete Temperatur and Regen). The data have been collected within the one-hour time interval.

In the project, ‘’Fahrradbruecke’’ column was nominated as the ‘’Total bike count’’ in the city regarding the explanation in the previous slide.

Overview on the Traffic Accidents Dataset

	UnfallID	UJAHR	UMONAT	Jahr-Monat	UWOCHENTAG	USTUNDE	UKATEGORIE	UART	UTYP1	ULICHTVERH	...	IstPKW	IstFuss	IstKrad	IstGkfz	IstSonstige	LINREFX	LINREFY
0	8,20E+18	2020	1	2020-01-01	6	14	2	9	1	0	...	1	0	0	0	0	510512,2678	5287973,343
1	8,20E+18	2020	1	2020-01-01	6	12	3	5	3	0	...	1	0	0	0	0	512502,3037	5279837,487
2	8,20E+18	2020	1	2020-01-01	5	15	3	0	7	0	...	0	0	0	0	1	513454,3717	5279340,436

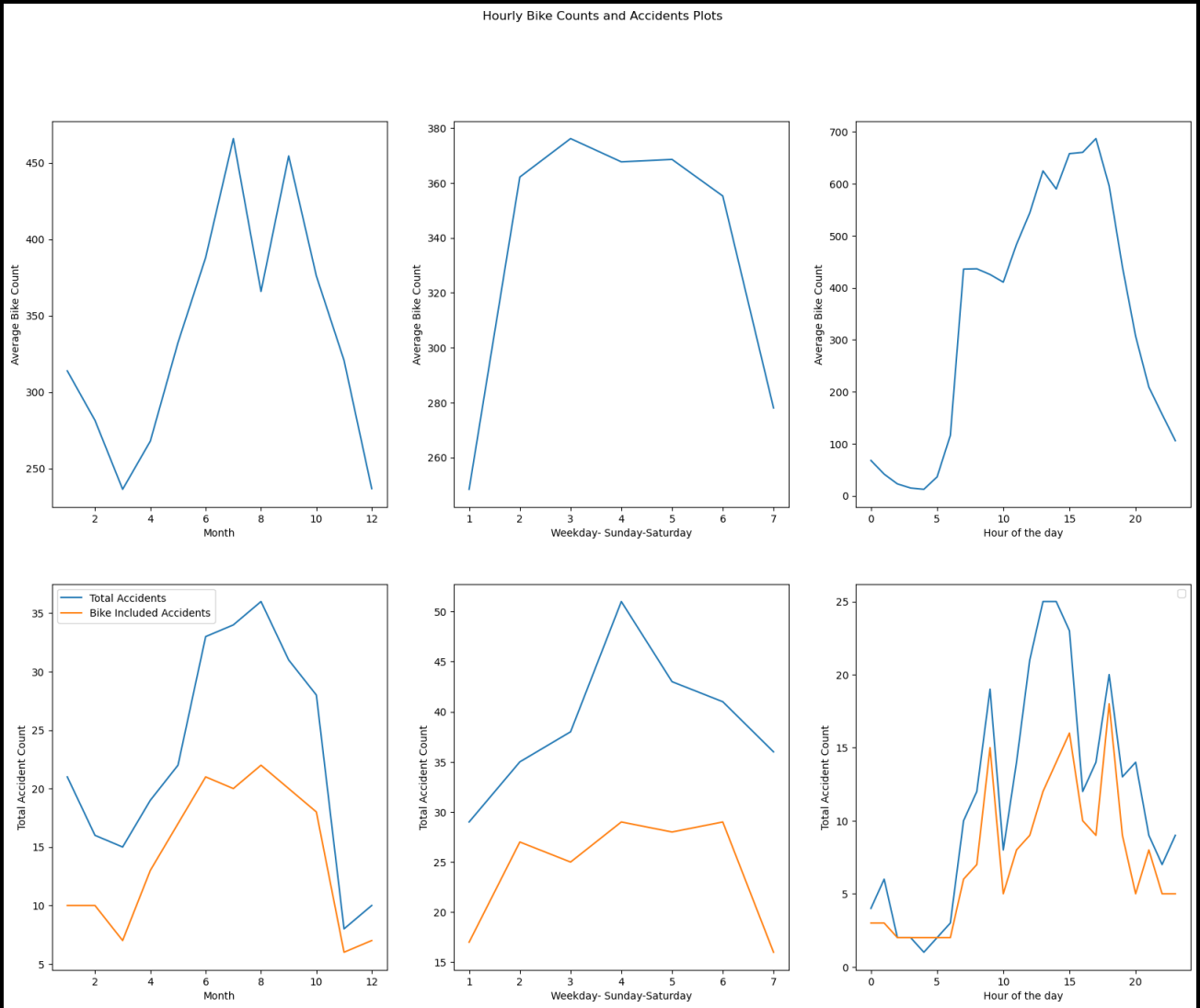
273 rows × 22 columns

The city of Konstanz, collected traffic accident data with the consideration of the various conditions such as the severity of the accident, a bike or motorbike was involved in the accident, the exact location of the accident, a pedestrian was involved in the accident, how was the lighting condition etc.

Possible questions:

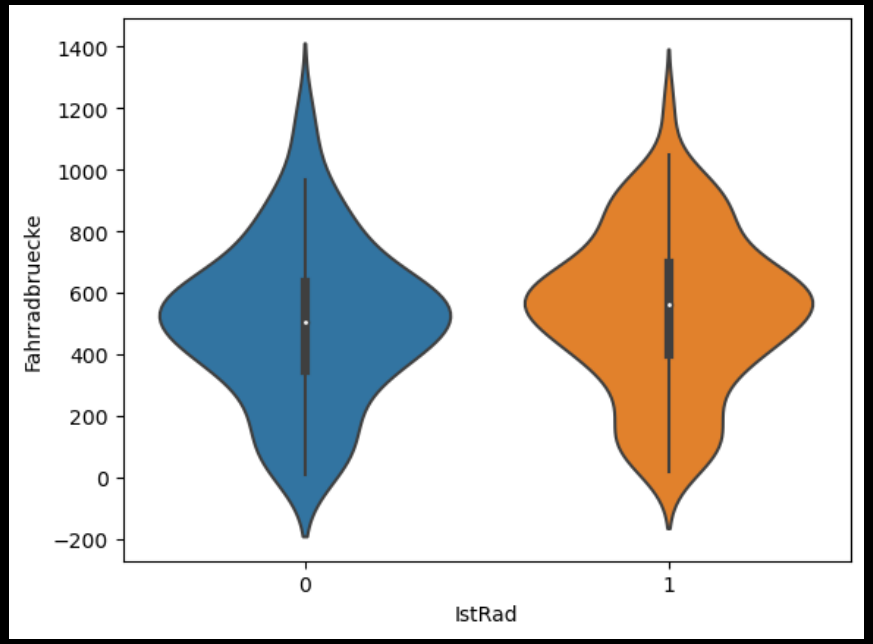
- How bike counts and traffic accident have changed subject to the date parameters (month, day of the week and hour of the day)?
- Is there a meaningful relationship between bike-involved accidents and average bike counts on flow?
- What really affects the bike counts and traffic accidents?
- What would be the best prediction tool to predict bike counts and bike involved-traffic accidents?
- What is the general distribution of traffic accidents over the city? Where are the tricky spots for the bike user?

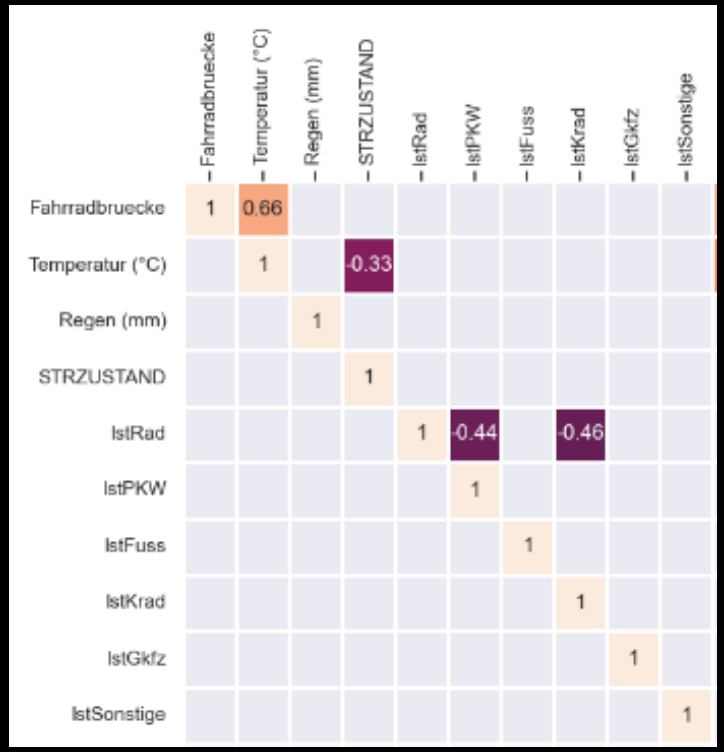
How bike counts and traffic accident have changed subject to the date parameters (month, day of the week and hour of the day)?



Regarding monthly analysis, July (Month 8) has the highest accident counts, however, in the summer, the total bike count of July has the lowest value during the summer season. On the weekly scale, the bike usage and bike included accident counts on the weekdays are way bigger than at the weekends. Lastly, at the hourly scale, we see the most bike active hours between 15-17 PM, on the other hand, the bike-included accidents drastically happened at 9 AM, 15 PM and 18 PM.

Is there a meaningful relationship between bike-involved accidents and average bike counts on flow?





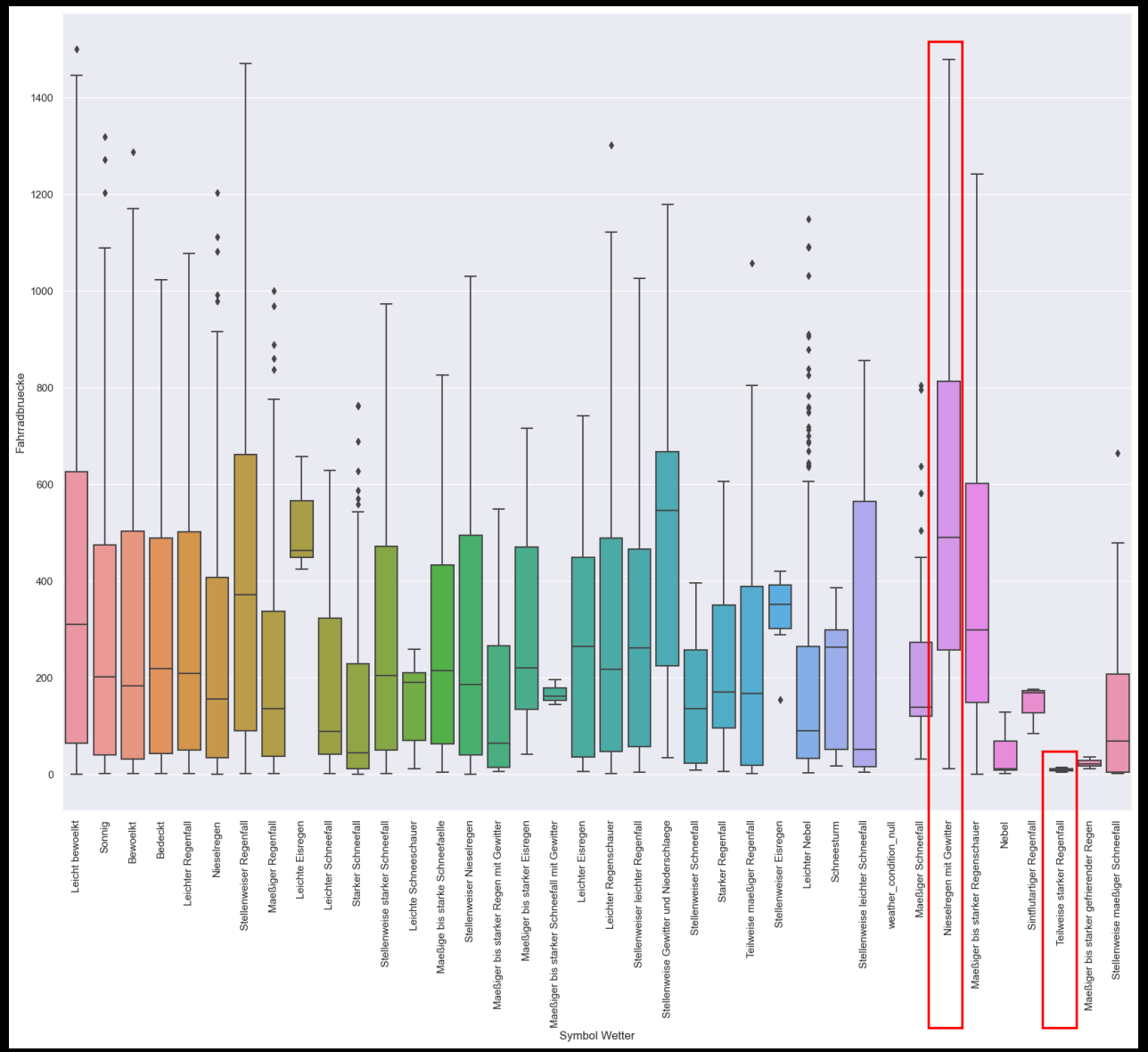
As we have seen on the violin plot, the data has been filtered by bike inclusion status and this parameter has been used for comparison parameters at the x-axis. Unexpectedly, both data (bike-included accidents and not included accidents) show almost the same median and even more or less the same distribution regarding to the bike count.
On the right, there is a little snippet from the correlation map of the merged dataset, again it shows that IstRad and Fahrradbruecke have no meaningful correlation. (Correlation matrix masked for the values over absolute .25)

What really affects the bike counts and traffic accidents?

According to this question, the following topics will be examined:

- The weather type on the bike count
- The feeling temperature on the bike count
- The categorical values on the traffic accidents

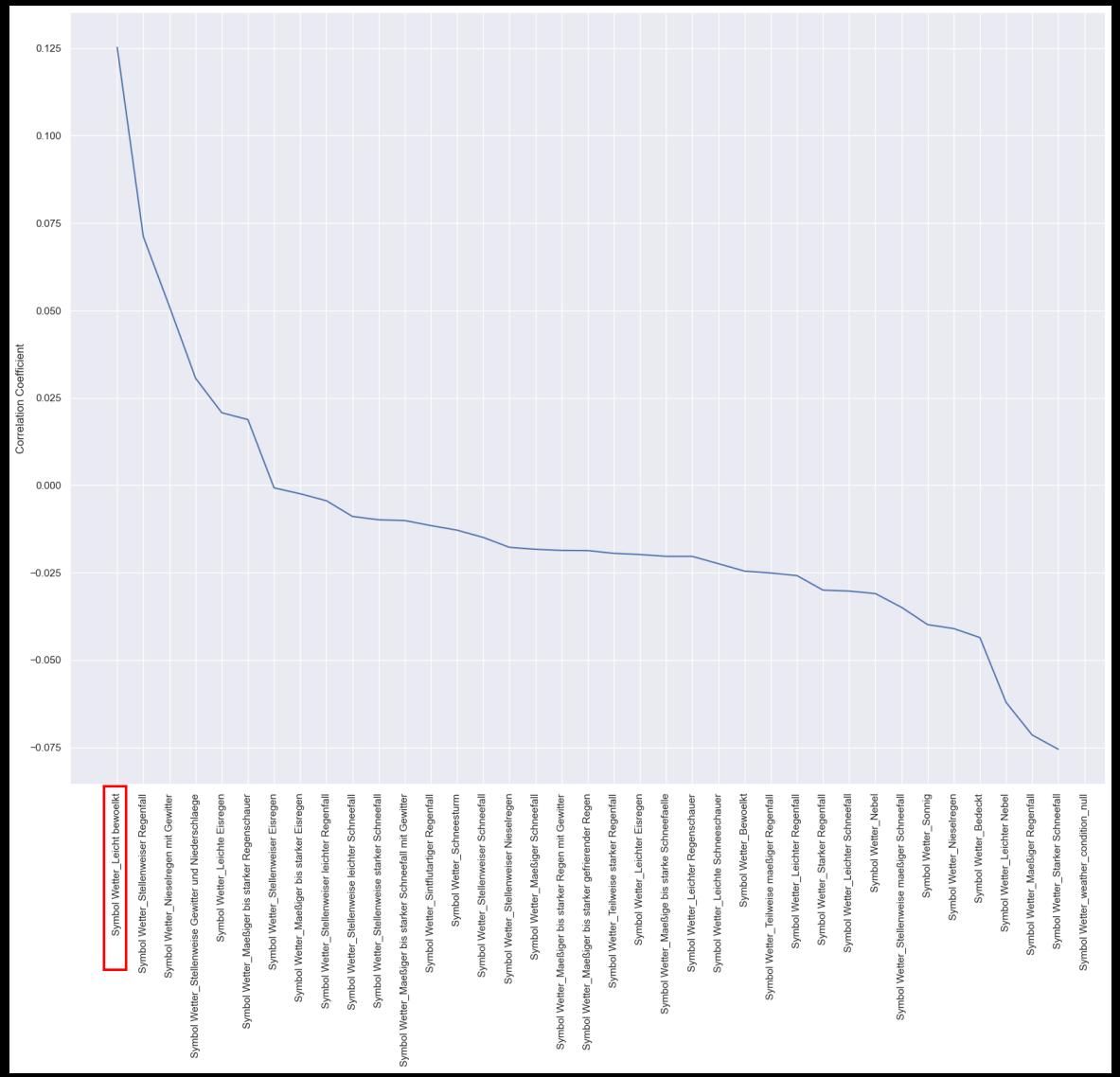
The weather type on the bike count



Weather Type	English Translation	Average Bike Count
Nieselregen mit Gewitter	Drizzle with thunderstorm	574.25641
Teilweise starker Regenfall	Partly heavy rain	9.666667

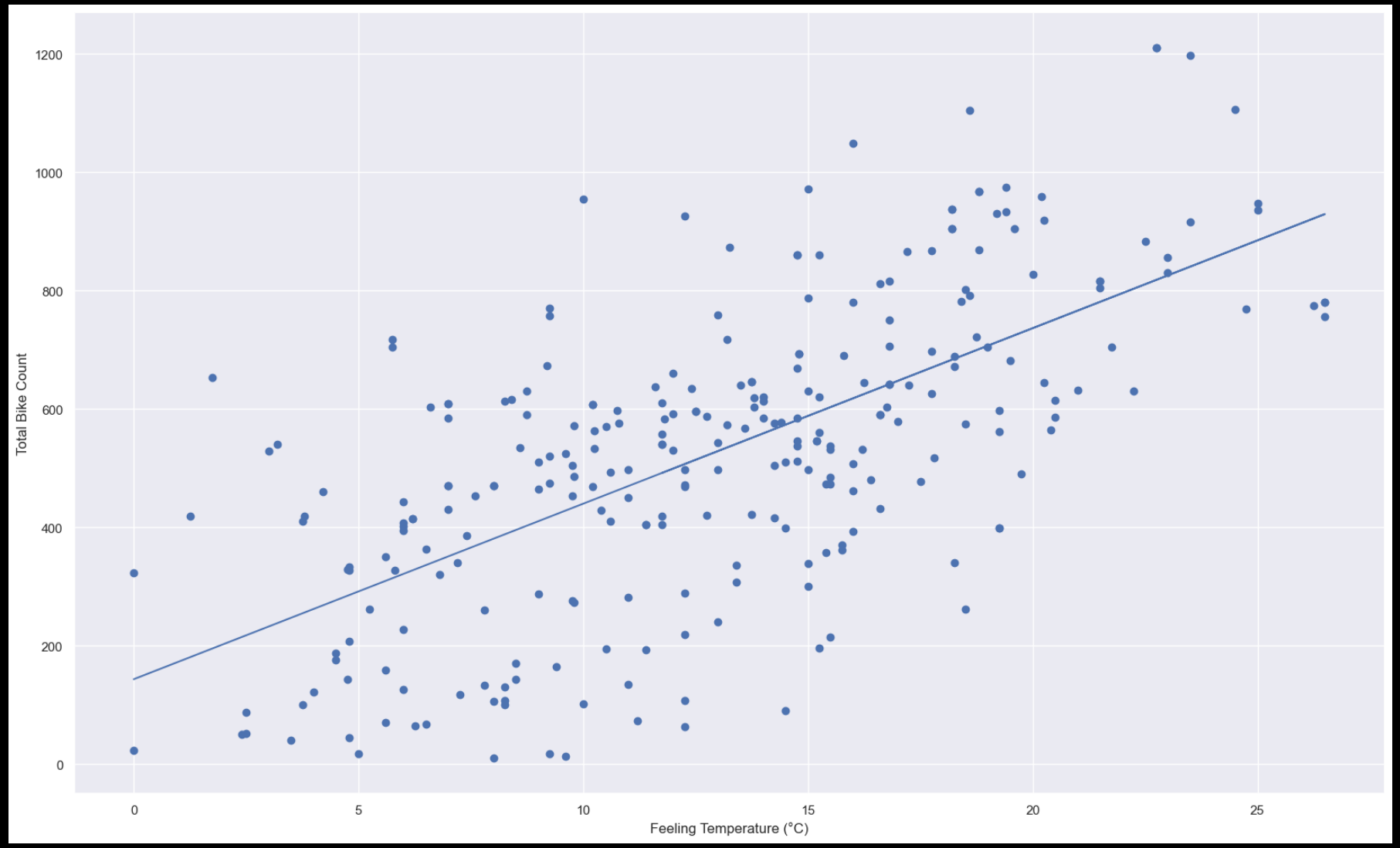
Interestingly "Nieselregen mit Gewitter" has the highest average bike count while the ‘Teilweise starker Regenfall’ leaded lowest.

The weather type on the bike count cnt.

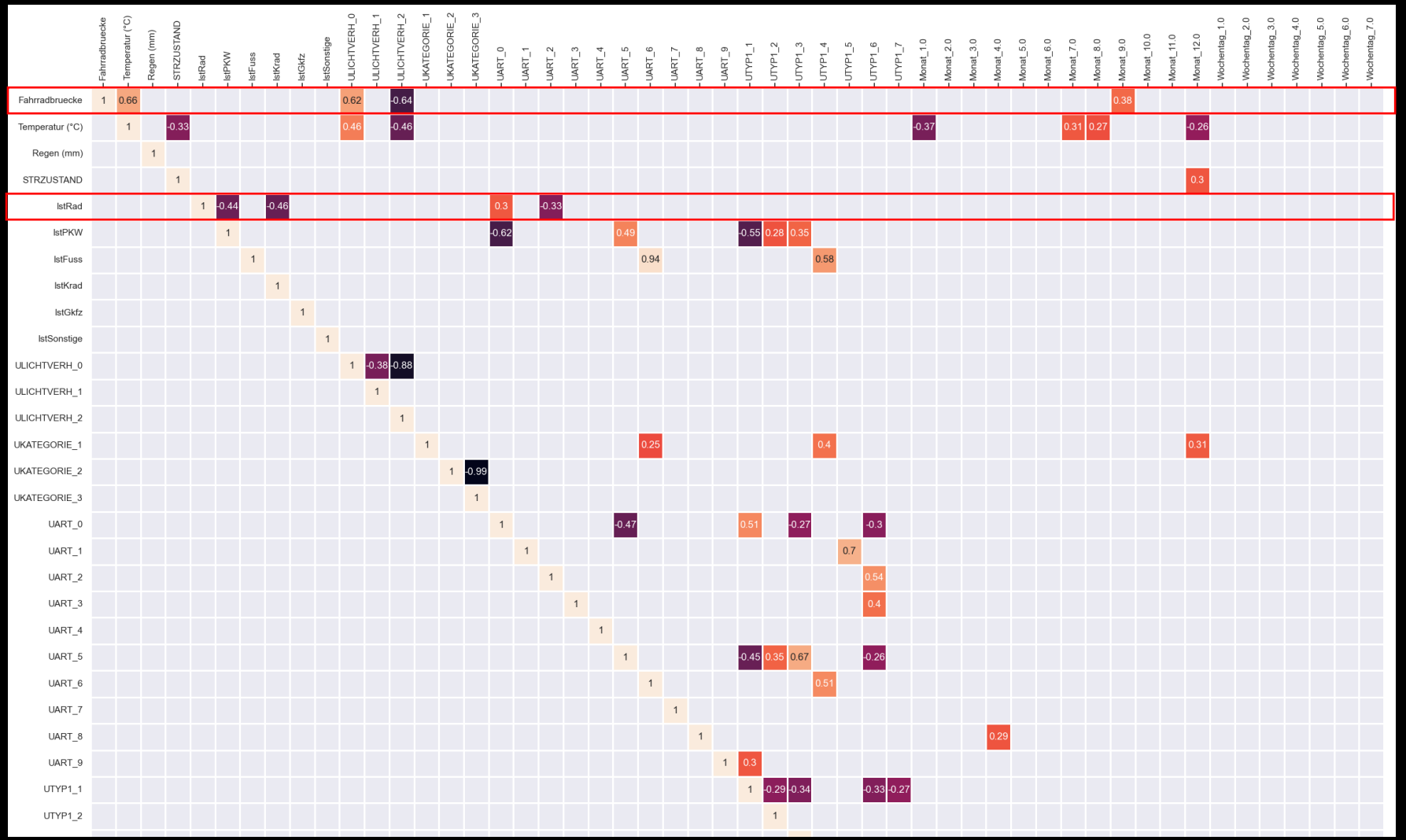


As we have seen on the map "Symbol Wetter" does not really affect the bike count. The highest correlation coefficient on the bike count has been seen on the "Symbol Wetter_Leicht bewoelkt" : 0.125.

The feeling temperature on the bike counts



The categorical values on the traffic accidents



As we see on the correlation matrix demonstrated, lighting condition (ULICHTVERH) especially 0 and 2 seem quite important effects on the total bike count but not on the accidents the bike involved "IstRad". However, "IstPKW'' (at least one passenger car was involved) and "IstKrad " (motocycle was involved) has prominent importance comparatively the other parameters.

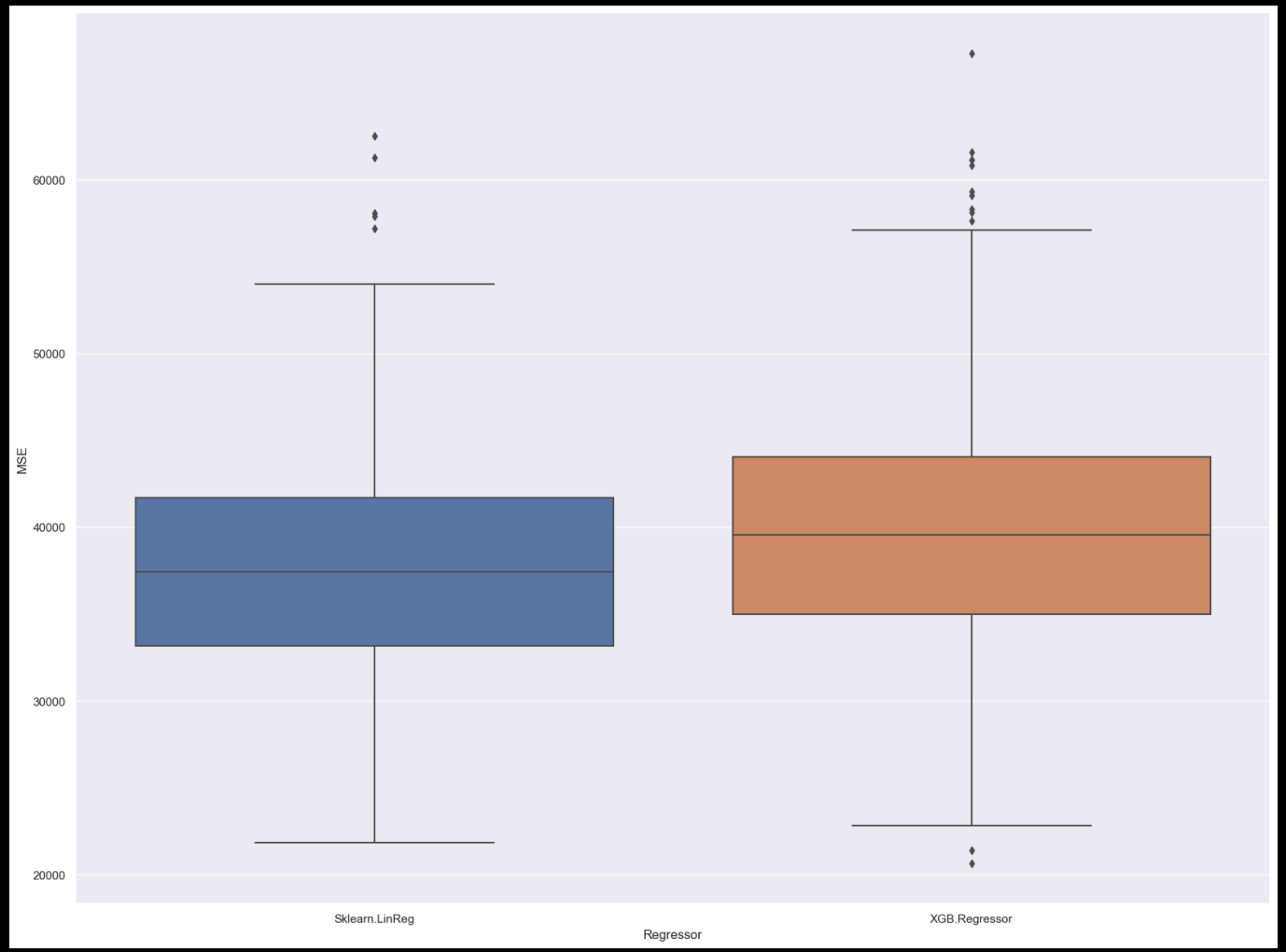
What would be the best prediction tool to predict bike counts and bike involved-traffic accidents?

As we already analyzed with a scatter plot, feeling temperature has a significant importance on the bike in the flow. But for finding the best prediction tool, we should compare at least two different tools. For the bike count prediction the following models were used:

- Sklearn.LinearRegression
- XGBoost.Regressor

The model qualities were evaluated by the MSE (mean squared error) metric.

Regressor Comparison



Both models were run 1000 times with the same dataset. And it clearly seen that sklearn.linreg slightly surpassed the Xgb.regressor for the prediction of the total bike count subject to the feeling temperature. In addition to this, it is a way better approach in terms of explainability.

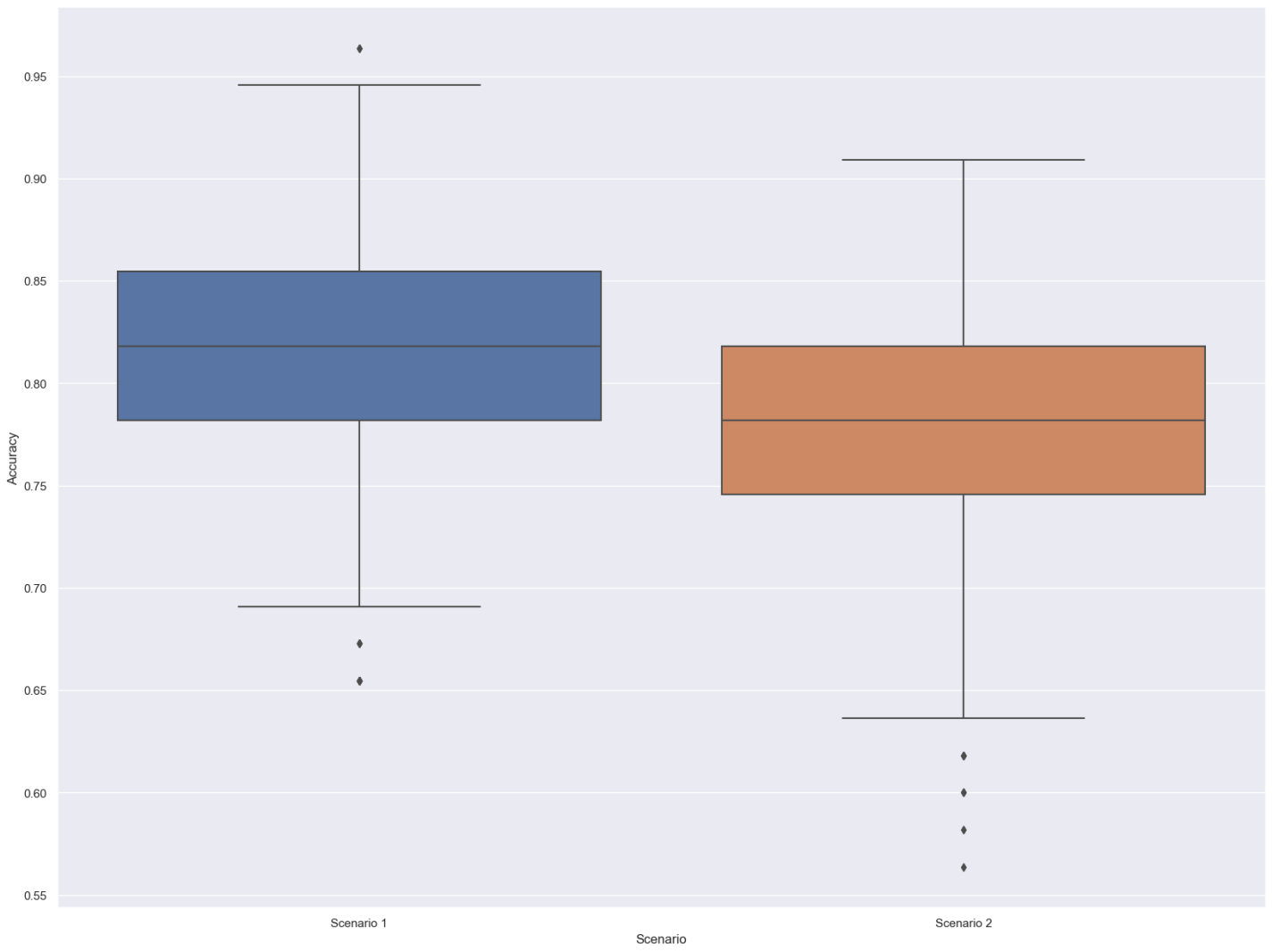
What would be the best prediction tool to predict bike counts and bike involved-traffic accidents? Cnt.

Even though the Sklearn regressor left behind the XGBoost regressor, XGBoost has pretty positive acknowledgement in data science especially for classification tasks. Therefore, we will use the XGBoost classifier as a prediction model. We will build the model under two different scenarios and compare the prediction accuracy:

- Scenario 1: With all features
- Scenario 2: Only selected features ['IstKrad','UART_0','UART_2','IstPKW']

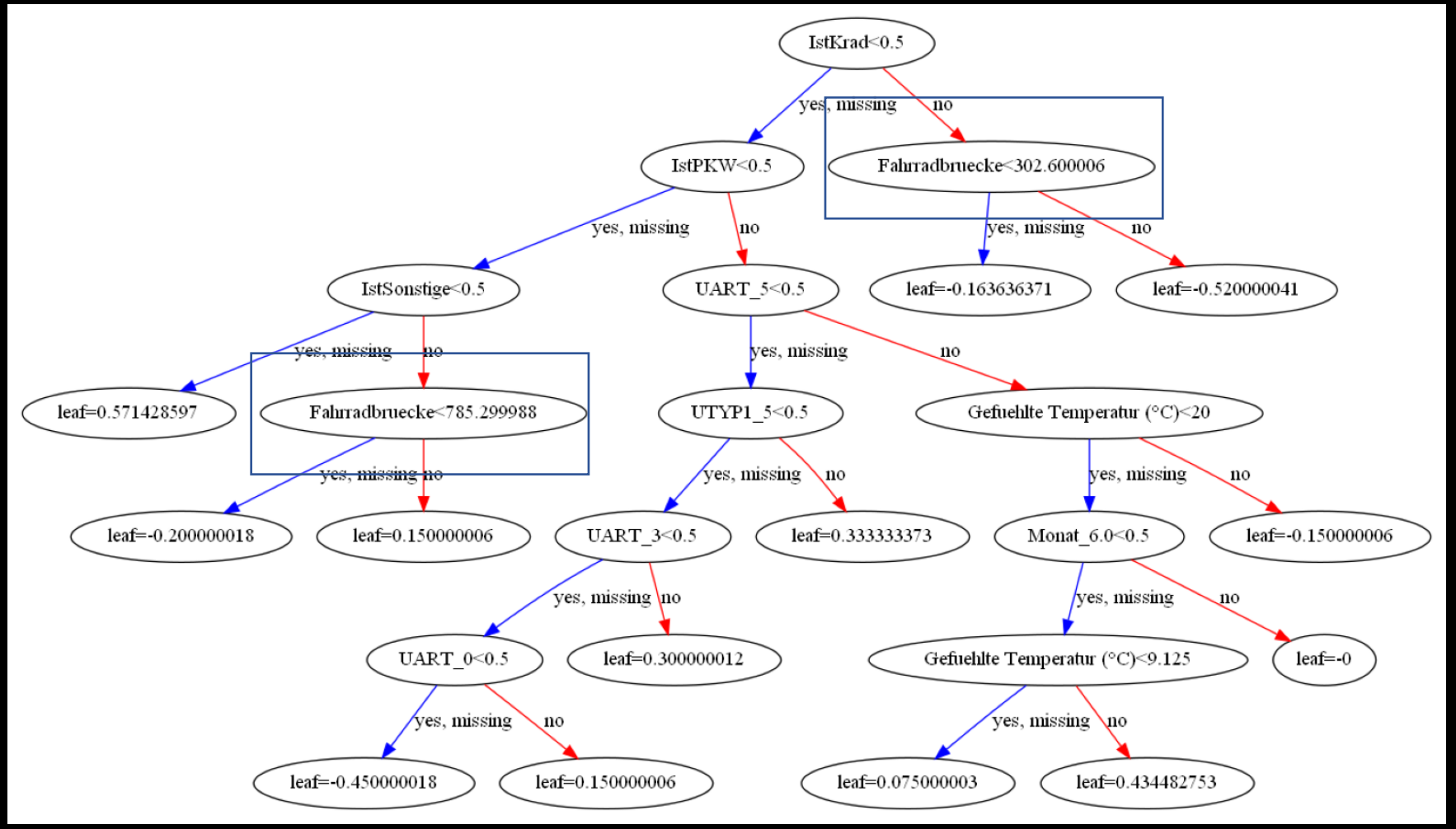
Two scenarios can lead different results at the every run because of the randomization at the data splitting phase. To decide which model is better we run both scenarios and collect MSE values of each and plot them on boxplot.

Scenario Comparison

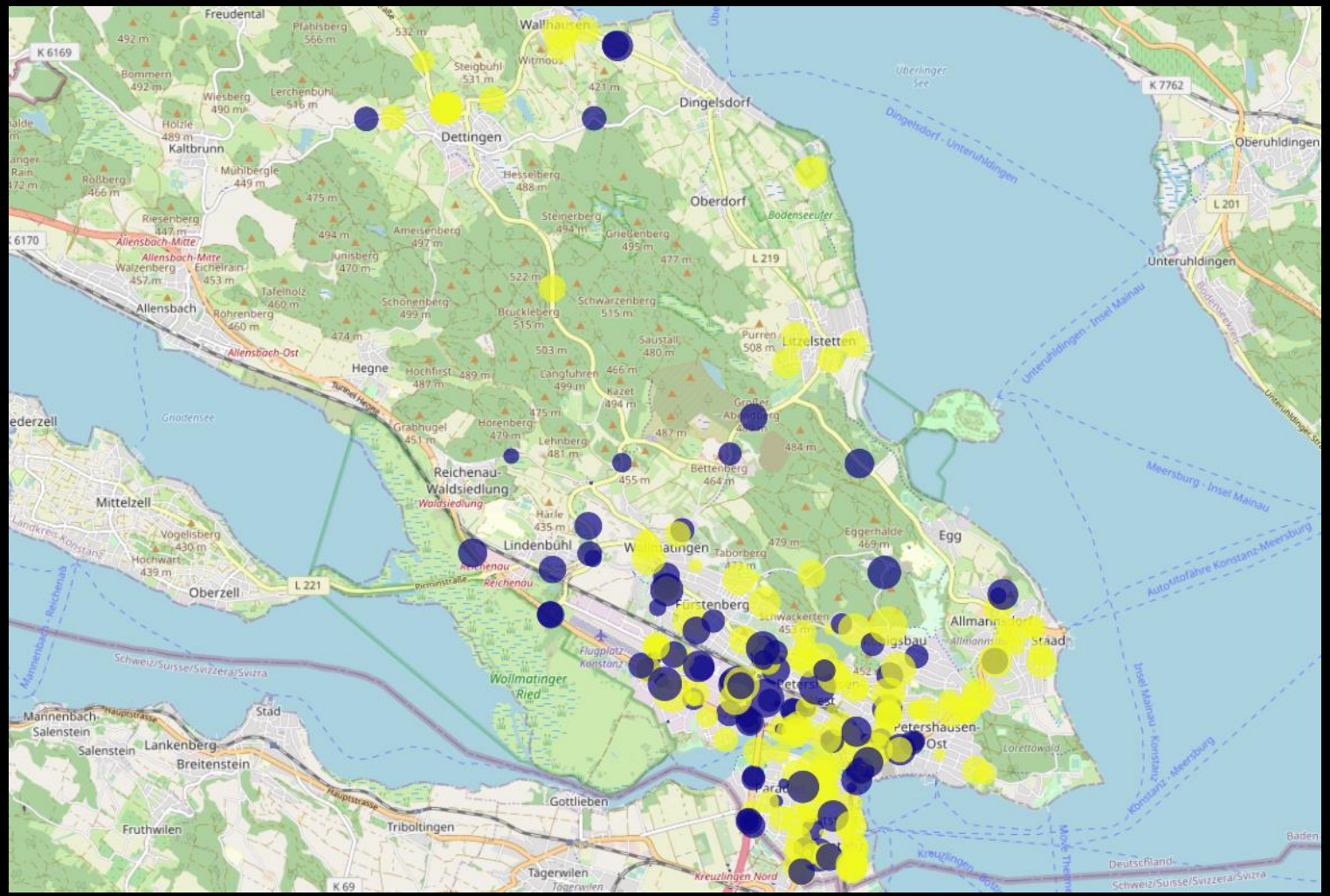


Again the both scenarios were run 1000 times and a boxplot obtained for the comparison. As the boxplot demonstrated Scenario 1 has almost 4% higher accuracy than Scenario 2 (in median). So we can conclude that for predicting the "IstRad" value, four features can give promising results but with all meaningful features we get better results in spite correlation matrix showing most of them as an uncorrelated feature.

Visualization of the XGBoost Decision Tree



What is the general distribution of traffic accidents over the city?

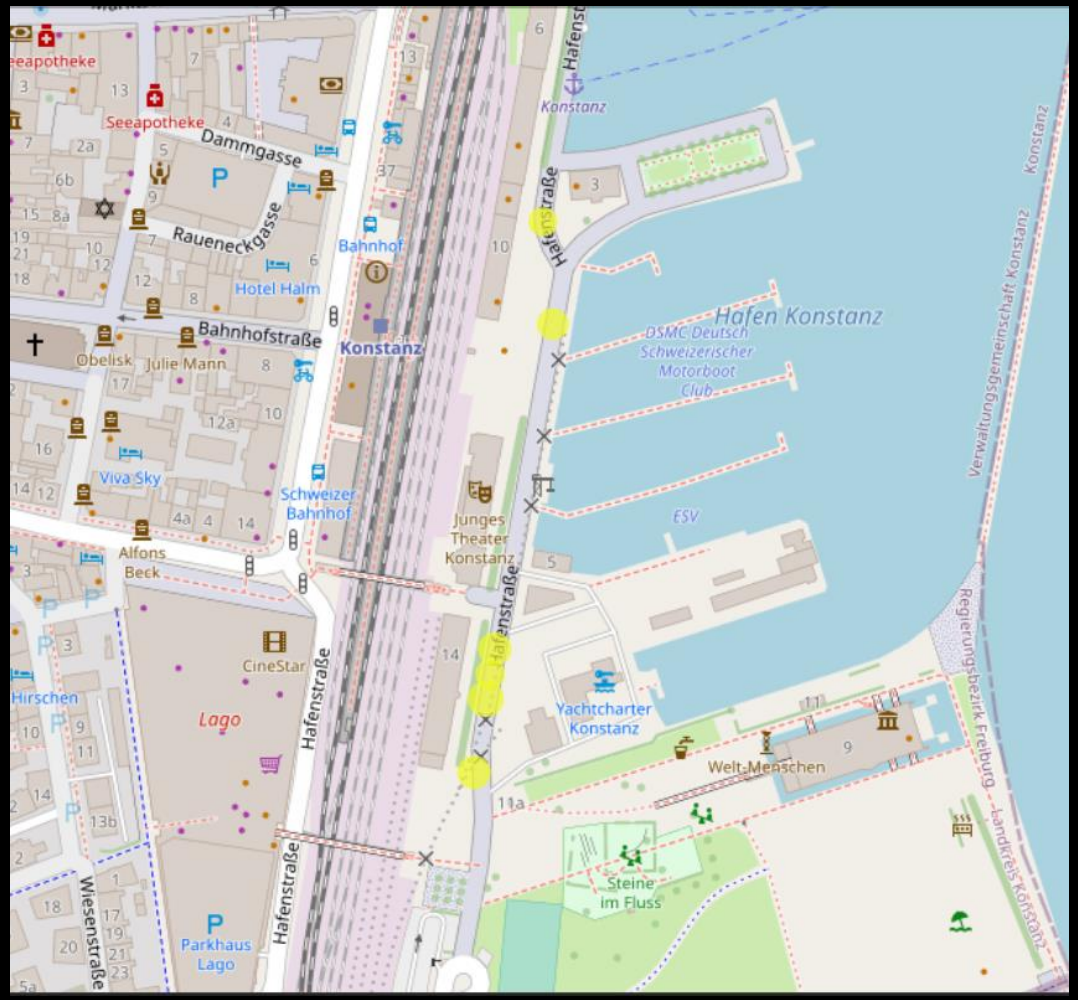


Yellow dots: Bike involved accidents
Blue dots: Other

The size of the blobs refer bike counts at that time

Where are the tricky spots for the bike user?

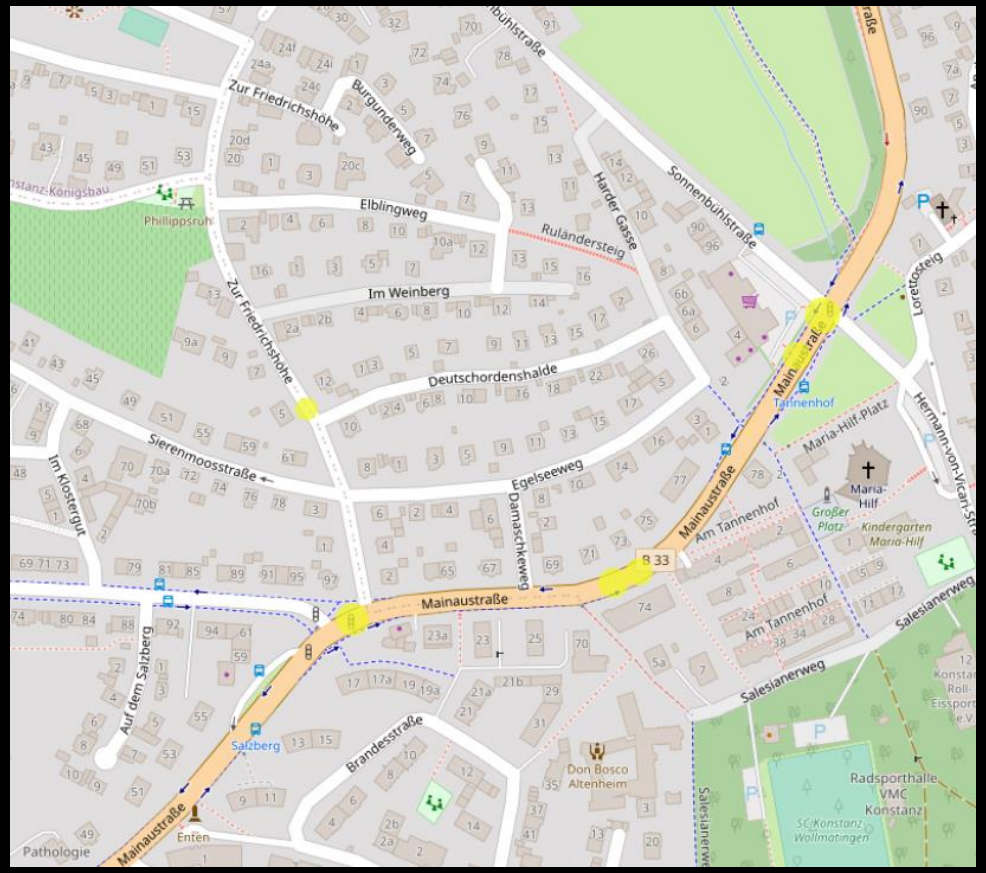
Hafenstraße by the Bay

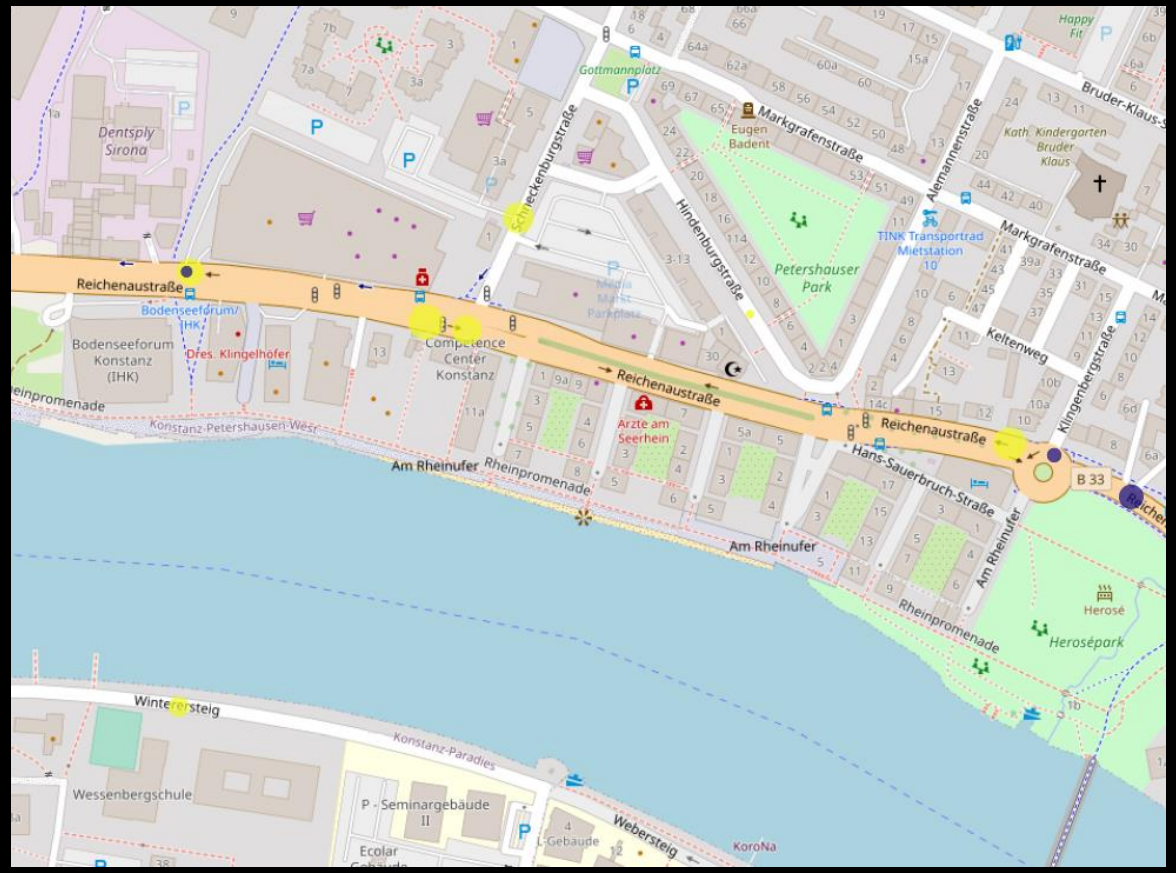


Where are the tricky spots for the bike user?

Bicycle Bridge Exit

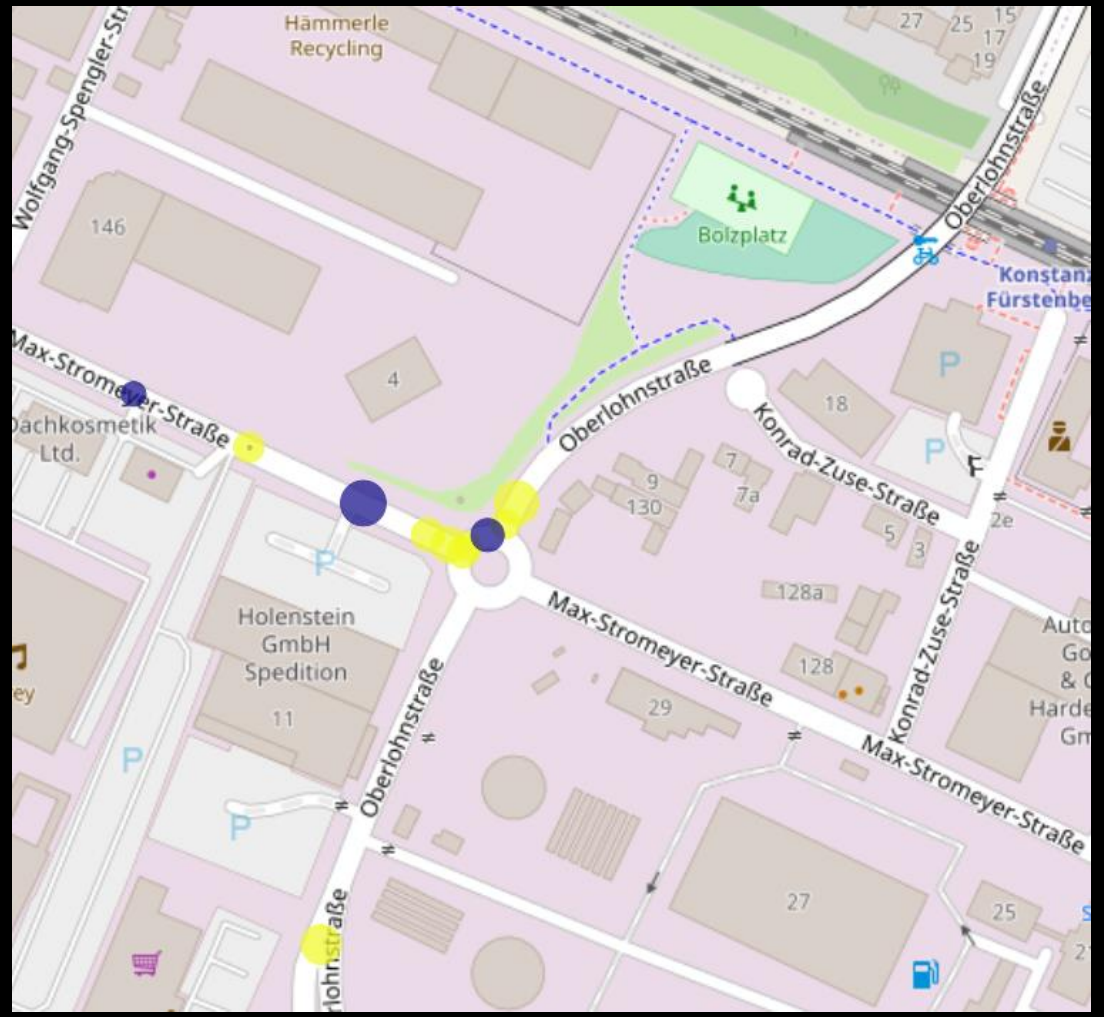
Highways


Zur Friedrichshöhe
Burgunderweg
Elblingweg
Ruländersteig
Harder Gasse
Sonnenbuhlstraße
Im Weinberg
Deutschordenshalde
Sierenmoosstraße
Egelseeweg
Damaschkeweg
Mainaustraße
Am Tannenhof
Maria-Hilf-Platz
Maria-Hilf
Großer Platz
Kindergarten Maria-Hilf
Tannenhof
Loretosteig
Hermann-von-Vicari-Straße
B 33
Salesianerweg
Im Klostergut
Auf dem Salzberg
Salzberg
Brandesstraße
Don Bosco Altenheim
Radsporthalle VMC Konstanz
Enten
Pathologie
Phillippsruh


Reichenaustraße
Bodenseeforum Konstanz (IHK)
Dres. Klingelhöfer
Competence Center Konstanz
Markgrafenstraße
Gottmannplatz
Eugen Badent
Hindenburgstraße
Schleckenburgstraße
Petershauser Park
TINK Transportrad Mietstation 10
Kath. Kindergarten Bruder Klaus
Alemannenstraße
Keltenweg
Klingenbergstraße
Hans-Sauerbruch-Straße
B 33
Konstanz-Petershausen-Weg
Am Rheinufer
Rheinpromenade
Ärzte am Seerhein
Herosé
Herosépark
Wintererrsteig
Konstanz-Paradies
Wessenbergschule
P - Seminargebäude II
Ecolar
Webersteig
KoroNa

Where are the tricky spots for the bike user?

Rotary Intersections



Conclusion

- There is no notable effect of the bike count in the flow and bike involved accidents. And there is no significant correlation.
- 9 AM, 15 PM and 18 PM are critical hours in the day for the bike users. And July is a bad month for all drivers.
- Weather types are not quite correlated with the bike count.
- There are some challenges spots inside the city that bike riders should take into account.
- For predicting bike counts, sklearn.linearregression whilst for the bike involved accidents xgboost classifier is the best method.