# Navigating Zurich

## A Comprehensive Analysis of Urban Traffic Dynamics

Project Presentation by

Pascal Sager, Luca Zhao, Weijia Zhong, Xiaohan Zhu

Image from wwf.panda.org

# Research Questions

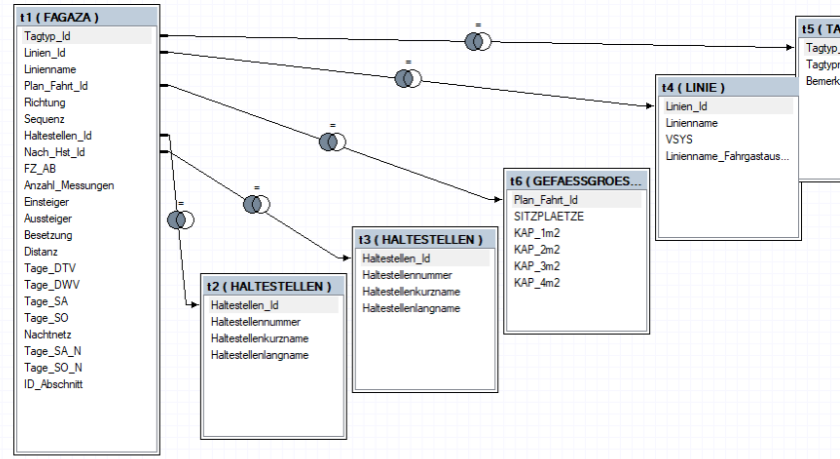
We analysed public transportation in the city of Zurich focused on:

- The **development** of Zurich's public transportation system over time
- The **utilization** intensity of Zurich's public transportation infrastructure
- The **spatiotemporal distribution** of passengers
- Analysis of the **interplay between diverse factors**, such as academic calendar and locations
- **Prediction of seat availability** on public transport

# Data

- **Passenger numbers boarding, aboard, alighting** for all VBZ traffic routes were sampled. For a given day type, and the average values of passengers were returned with information of the number of days in the day types
  - 1M+ entries
  - 39 features (columns)

- 6 External tables with information on vehicle type, capacity, day type, etc.
- Extended original dataset with historical data and GPS coordinates

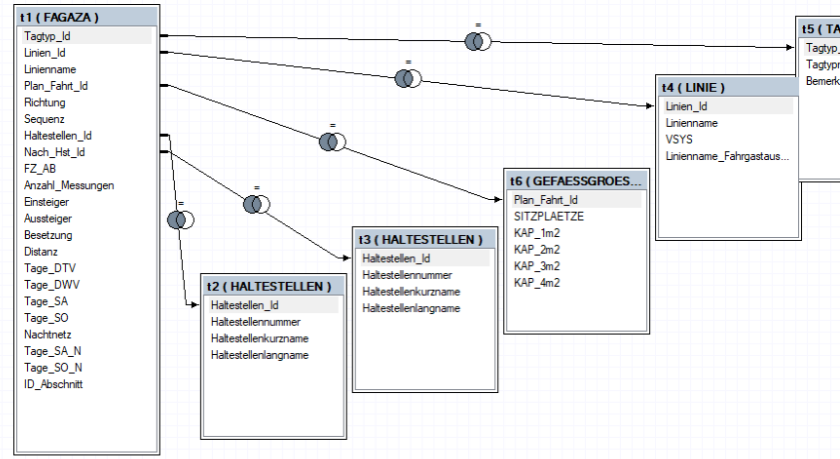



# Data Quality

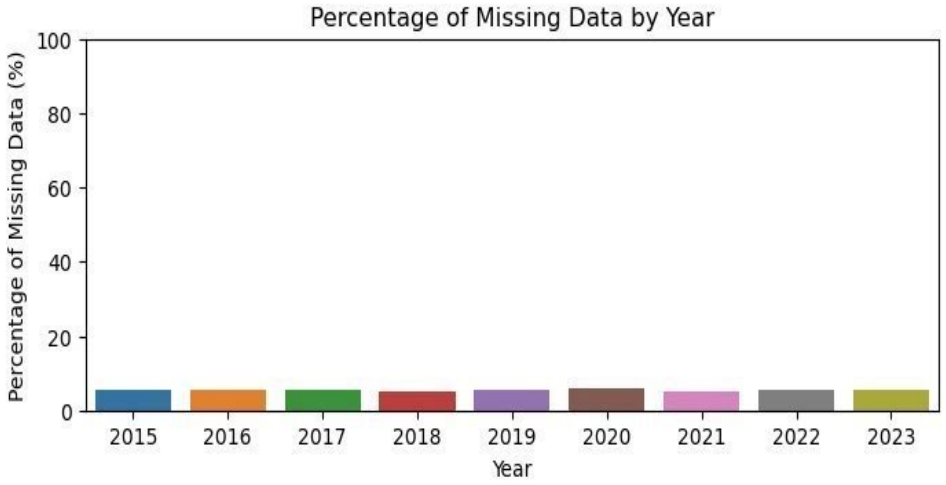
The overall quality of the data is decent, there are some **inconsistencies across different years**.

- Missing values are present
- Departure time shifted for 4 hours
- Corrupted encoding type
- Inconsistent delimiter

...

# Data Cleaning




- **Missing data**

  We simply removed the rows because those entries are final stops without passengers onboard.

- **Incomplete data about the GPS coordinates**

  We used the library GeoPy to retrieve the coordinates and manually correct generated data.

- **Inconsistencies across different datasets**

  We detect the corrupted characters and replace them with corrected ones.

- **Shifted departure time**

  We apply the function that for each row to take the remainder of division by 24.

# Development of Zurich's Public Transport System Over Time

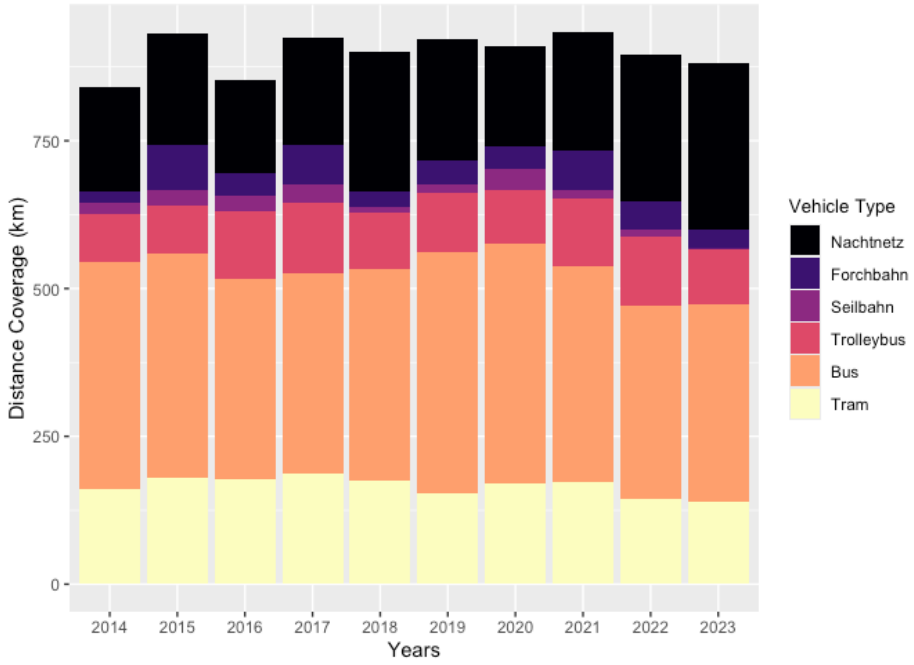
# The Spatial Coverage (km) of VBZ over the Years

**Method:**

- Filter the dataset for unique stops, sum up the distance in between.

**Findings:**

- Only some small changes happened across year 2014 to 2023. Possibly due to VBZ minor adjustments.




Zimmerberg

Lines 161, 165, 184 and 185

New with electric buses

Modern electric articulated buses are used on the lines on the Unterer Zimmerberg.

Line 161 and 163

Kilchberg, Community Centre

The stop Kilchberg ZH, Kreuzstrasse will be renamed Kilchberg ZH, Gemeindehaus.

Line 162 and 184

Kilchberg, Asp

The stop Kilchberg, Spital will be renamed Kilchberg, Asp, as the See-Spital can no longer be found at this location.

Line 165

More trips on weekends

Line 165 will be increased to four trips per hour (every 10/20 minutes) between Bürkliplatz and Kilchberg, Lindt & Sprüngli on Saturdays and Sundays from 10 a.m. to 7 p.m. The Lindt & Sprüngli stop will be moved about 160 meters towards Zurich on Seestrasse. The former Schooren stop will be reactivated for this purpose.

Line 184

Early morning to Wollishofen

During the week, the first bus leaves Adliswil for Wollishofen half an hour earlier, so you arrive at Wollishofen station at 5:53 a.m.

# Passenger Volume Change Across the Years (mio.)

**Method:**

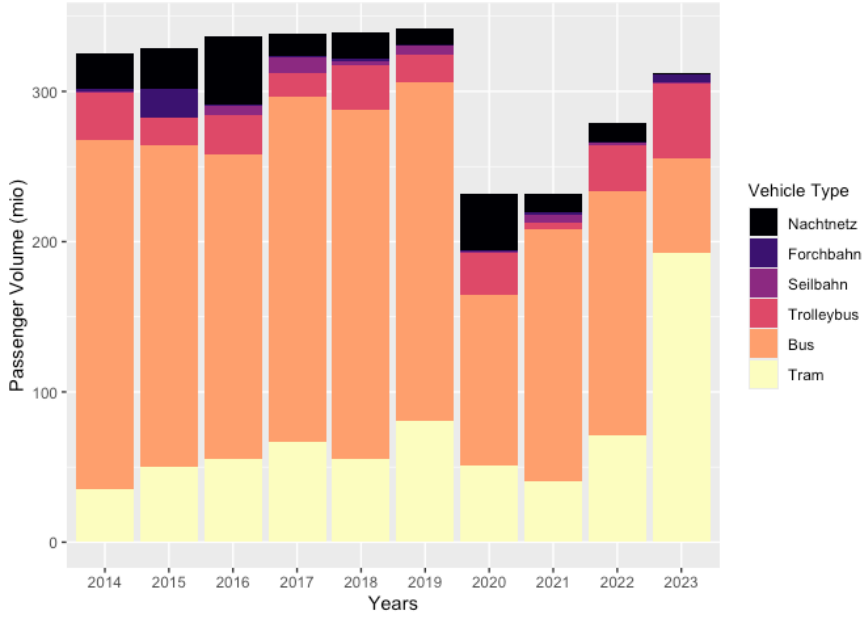
- The number of passengers aboard for

  (unique rides per day * number of days).

**Findings:**

- Significant drop during COVID-19.
- Slowly gets back 2022-2023 but still not fully to the level pre-COVID.
- A major shift of passenger volume from bus to tram during 2022-2023, why?

Web search only shows one news about VBZ changing the frequency of many bus lines due to staff shortage




**What tram and bus services in Zurich are being cut?**

According to officials, tram line 15 will remain suspended until further notice and tram line 17 will only run until 8.30pm on Monday to Saturday. Line 17 will not run at all on Sunday. For passengers looking to travel to Werdholzli, line 13 will be extended to make up for the lost connection.

The reduction in service will also affect a number of bus routes around the city:

- Line 32 will only run every 7.5 minutes from January.
- Lines 40, 61 and 62 will only run every 15 minutes.
- Line 66 will only run services between Morgental and Neubühl every 15 minutes.
- Line 89 will end at Sihlcity after 9pm on Monday through Saturday, and at all times on Sunday.

# Passenger Volume Change in 24-hours of a Day (2023)

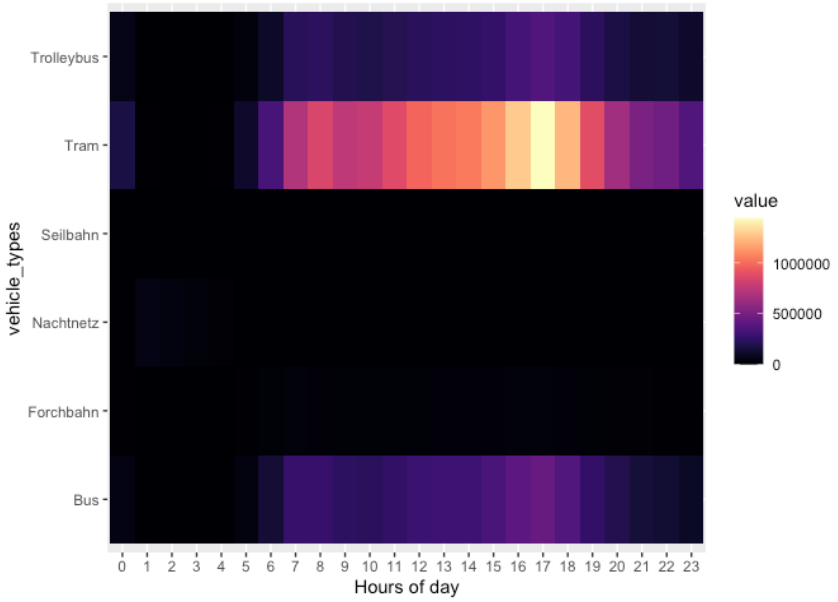
**Method:**

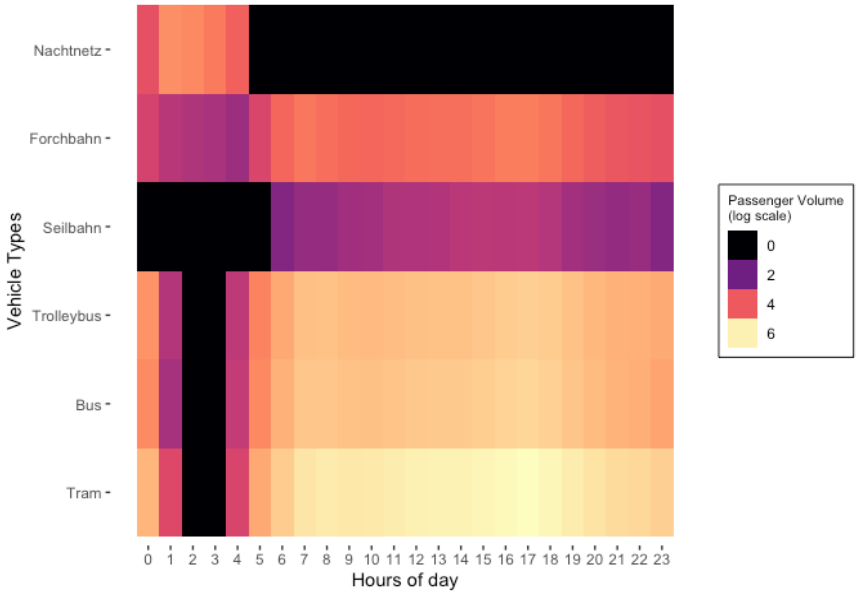
Total passenger numbers were computed at 1 hour interval for 24 hours in a day, w/w.o log scale

**Findings:**

Majority of public transportation volume fall on the tram system (same as previous plot).

Busy hours: 7-8 in the morning and16-18 in the evening, corresponding to commuting needs.



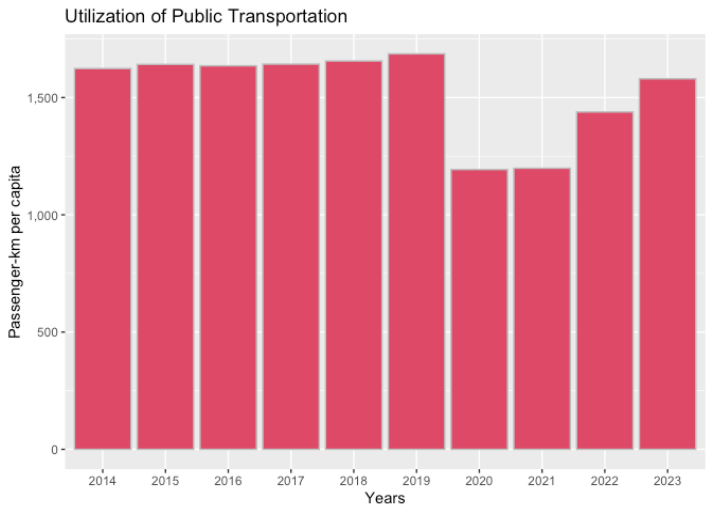
# Utilization of Urban Transportation in Zurich (per capita)

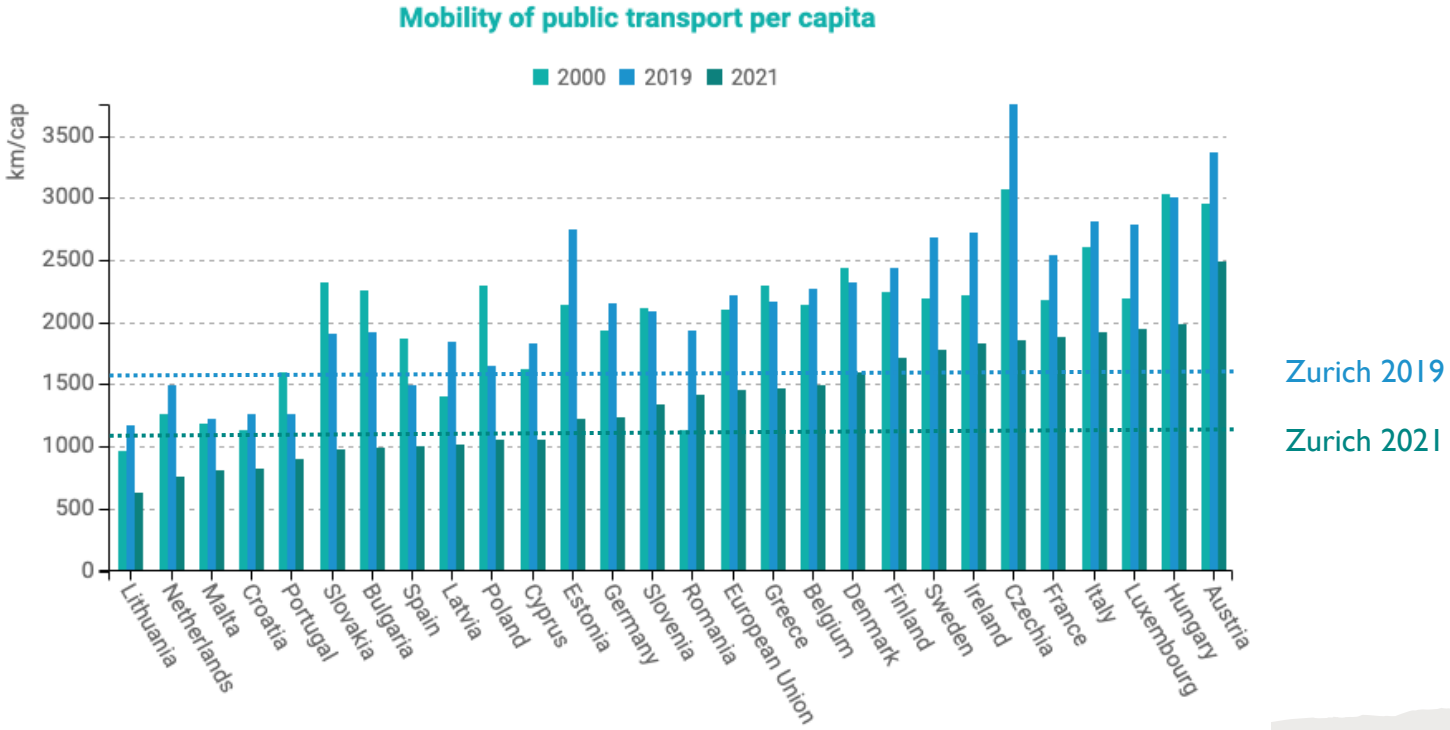
**Method:**

Passenger-Kilometre (pkm) per capita computed as

(unique rides per days * number of days * distance travelled) / 443,037

443,037 = number of inhabitants in Zurich city, 2023




Datasource: Odyssee-Mure
https://www.odyssee-mure.eu/publications/efficiency-by-sector/transport/passenger-mobility-per-capita.html

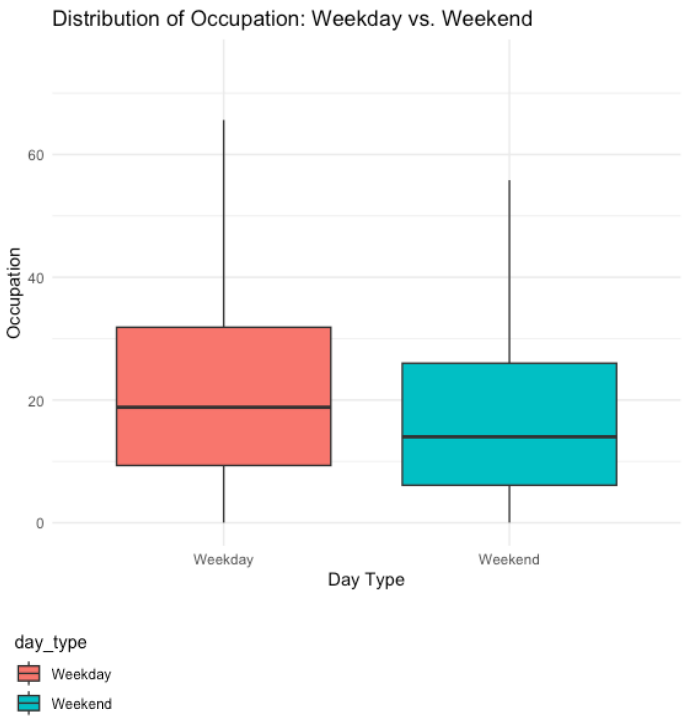
# Factor Analysis on Average Passenger Numbers

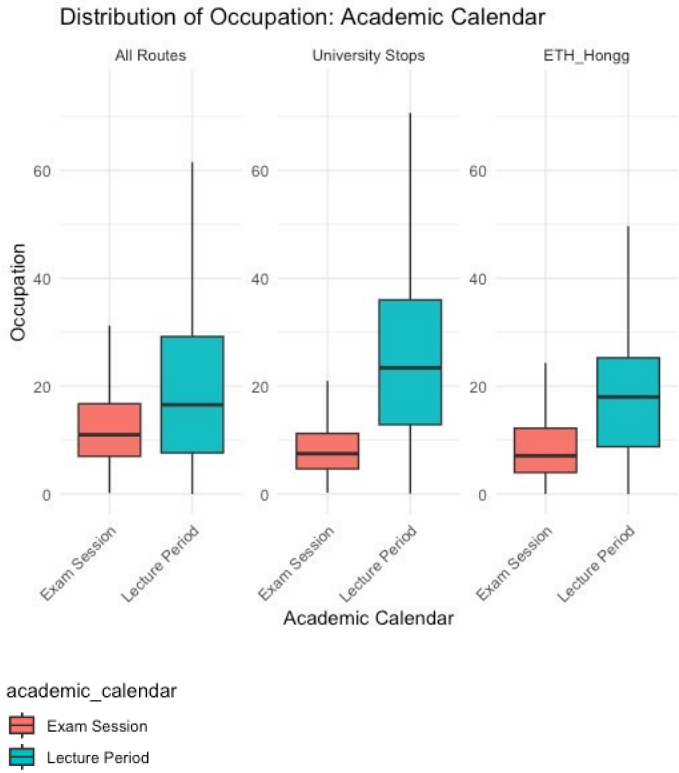
**Method:**

- Dataset group by factors then compute the mean of passenger numbers
- Distinction on stops in vicinity to universities

**Findings:**

- More passengers on average on weekdays, likely due to commuting needs
- Significant drop of passenger numbers in gross traffic might be due to seasonal factors such as extreme weather conditions



# Spatiotemporal Distribution of Passengers

# Geospatial Analysis

**Method**

- Merge data with GPS coordinates
- Aggregate data per hour
- Draw arrows on top of the map of Zurich
- Color arrows according to seat occupation


**Findings**

- Weekday: A lot of traffic around 8 am and 5 pm towards / away from the main station
- Irchel: Students arrive ~8 am and leave ~6 pm
- Weekend traffic constantly increases until a peak is reached around ~5 pm

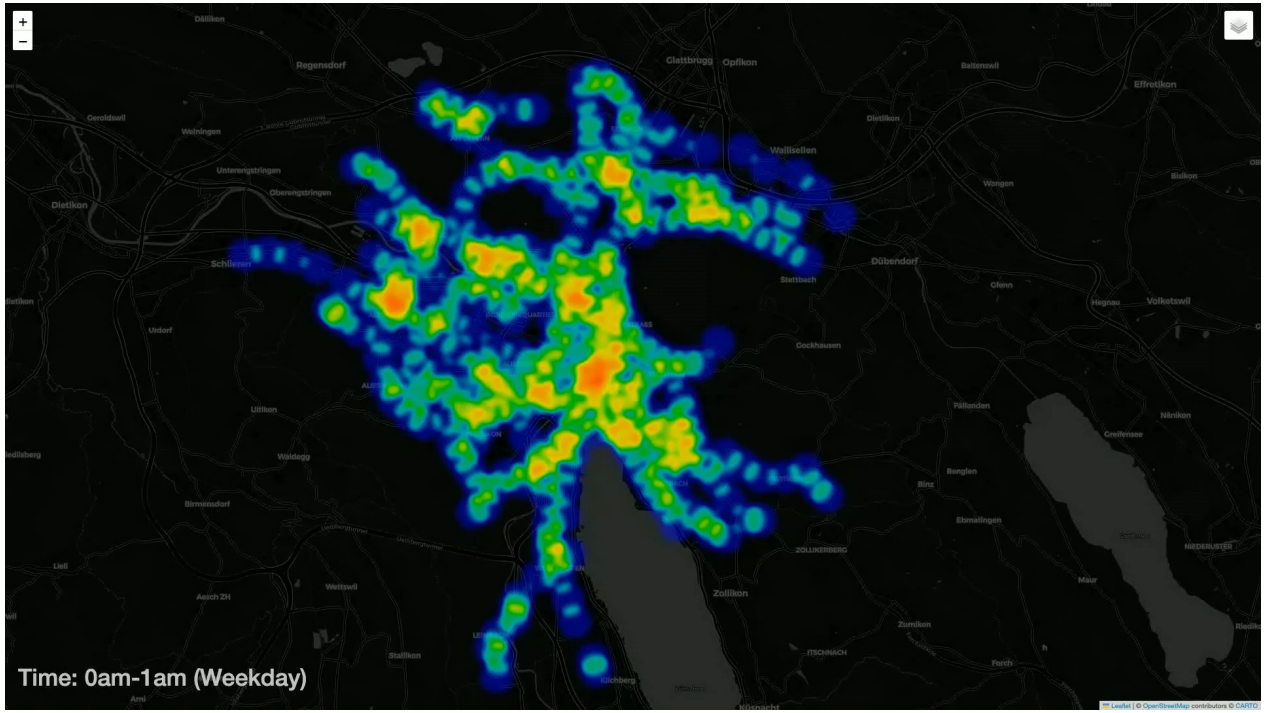
# Geospatial Analysis

**Method**

- Merge data with GPS coordinates
- Aggregate data per hour
- Sum passengers boarding and alighting at stops
- Plot heatmap on top of city of Zurich

**Findings**

- Busy areas: Main station, Altstetten, Paradeplatz, Bellevue



# Prediction on Seat Availability with Machine Learning

# Data Cleaning and Preparation

**Method:**

- Adjust inconsistencies between datasets
- Drop non-predictor columns and remove rows with missing data
- Transform categorical columns to numerical/boolean
- Incorporate geological coordinates to stops
- Normalization on variables

[188]:

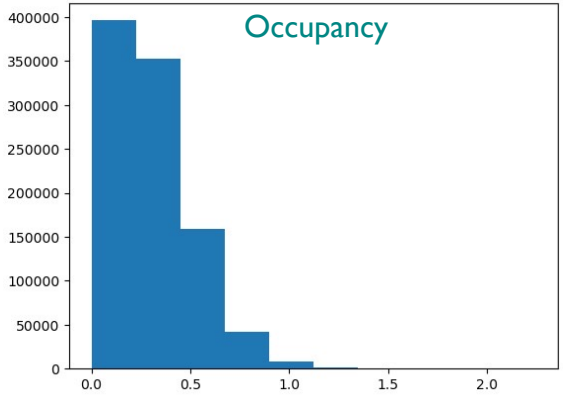
| | Time | Besetzung | Tage_SA | Tage_SO | Nachtnetz | Tage_SA_N | Tage_SO_N | Occupancy | Freeseats | GPS_Latitude | GPS_Longitude | Weekday | Richtung_1 | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 21.5 | -0.913107 | 0 | 0 | 0 | 0 | 0 | -0.878071 | -0.715621 | 0.474089 | -0.907114 | 1 | False | |
| 1 | 21.5 | -0.935909 | 0 | 0 | 0 | 0 | 0 | -0.936976 | -0.698837 | 0.474089 | -0.907114 | 1 | False | |
| 2 | 22.0 | -0.919992 | 0 | 0 | 0 | 0 | 0 | -0.895856 | -0.710554 | 0.474089 | -0.907114 | 1 | False | |
| 3 | 22.0 | -0.922084 | 0 | 0 | 0 | 0 | 0 | -0.901261 | -0.709014 | 0.474089 | -0.907114 | 1 | False | |
| 4 | 22.5 | -0.951654 | 0 | 0 | 0 | 0 | 0 | -0.977654 | -0.687246 | 0.474089 | -0.907114 | 1 | False | |
| 5 | 22.5 | -0.985617 | 0 | 0 | 0 | 0 | 0 | -1.065392 | -0.662246 | 0.474089 | -0.907114 | 1 | False | |
| 6 | 23.0 | -1.006237 | 0 | 0 | 0 | 0 | 0 | -1.118662 | -0.647067 | 0.474089 | -0.907114 | 1 | False | |
| 7 | 20.5 | -0.902409 | 1 | 0 | 0 | 0 | 0 | -0.850434 | -0.723496 | 0.474089 | -0.907114 | 0 | True | |
| 8 | 21.0 | -0.941041 | 1 | 0 | 0 | 0 | 0 | -0.950236 | -0.695059 | 0.474089 | -0.907114 | 0 | True | |
| 9 | 21.0 | -0.916277 | 1 | 0 | 0 | 0 | 0 | -0.886260 | -0.713288 | 0.474089 | -0.907114 | 0 | True | |
| 10 | 21.5 | -0.949956 | 1 | 0 | 0 | 0 | 0 | -0.973267 | -0.688496 | 0.474089 | -0.907114 | 0 | True | |

# Initial Attempt with Linear Regression and Random Forest

**Method:**

- 2022 data as the training set and 2023 data as the test set
- Relevant feature variables = Time, Vehicle type, GPS, Day type, all normalized

  Predictor variable = Occupancy [%], not normalized




- Use GridSearchCV for hyperparameter tuning
- MAPE, MAE, and MSE together with MAE of the median predictor were computed on the prediction of the test set for evaluation

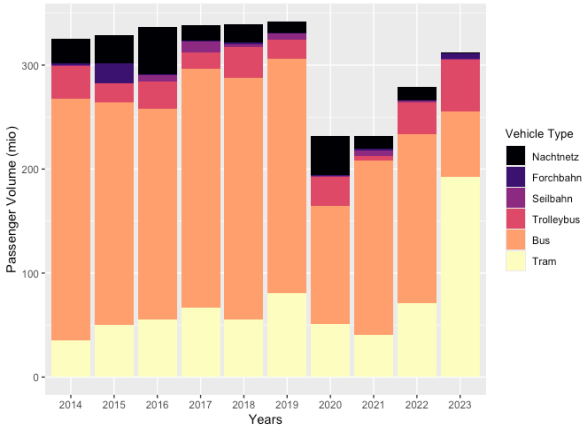
| Model | MAE [n passengers] | MSE [n passengers] |
|---|---|---|
| Linear Regression | 12.8 | 281.5 |
| Random Forest | 11.9 | 244.2 |
| **MAE of the median predictor: 11.8** | | |

# Troubleshooting the Prediction

**Method:**

- 2017 data as the training set and 2019 data as the test set for similarity in structure
- Relevant feature variables = Time, Vehicle type, GPS, Day type, all normalized

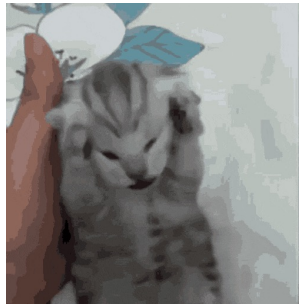
  Predictor variable = Occupancy [%], not normalized -> normalized




- Try Support vector regression (SVR) for better performance on non-linear patterns

  -> never finishes running, reduce data to tram only and use PCA for dimension reduction

| Model | MAE [n passengers] | MSE [n passengers] |
|---|---|---|
| Linear Regression | 23.1 | 880.9 |
| Random Forest | 18.9 | 680.7 |
| SVR | ~21.0 | ~800.0 |
| **MAE of the median predictor: 35.2** | | |



**Performance got worse -> Does not work**

# Prediction: Deep Learning

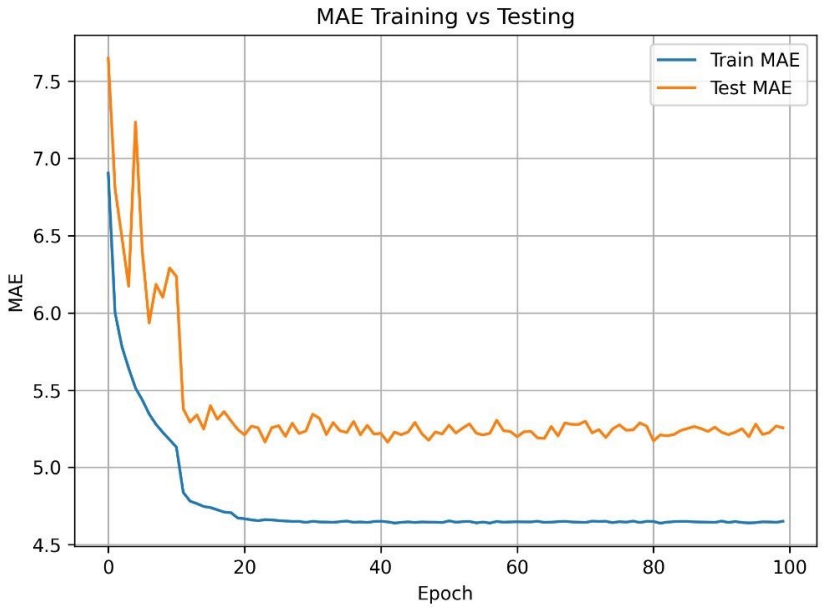
**Method:**

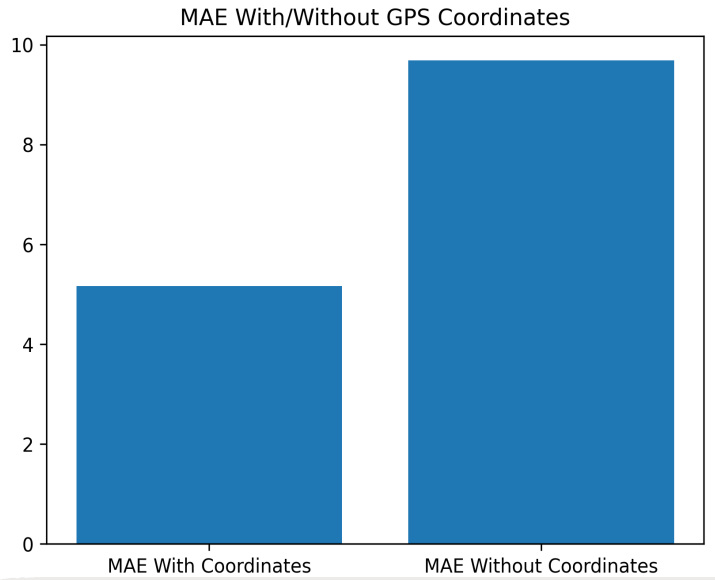
- Define the dataset and the model (fully connected layers with batch normalization and dropout).
- Train the model with and without adding GPS coordinates.
- The loss function used is Mean Squared Error (MSE) to heavily penalize larger errors.
- Adam optimizer with a learning rate of 0.01 to accelerate the convergence.
- Reduces learning rate by a factor of 0.1 if the loss does not improve for three epochs.

**Findings:**

- We obtain far better results with the deep learning model; with added GPS coordinates, the results are even better.

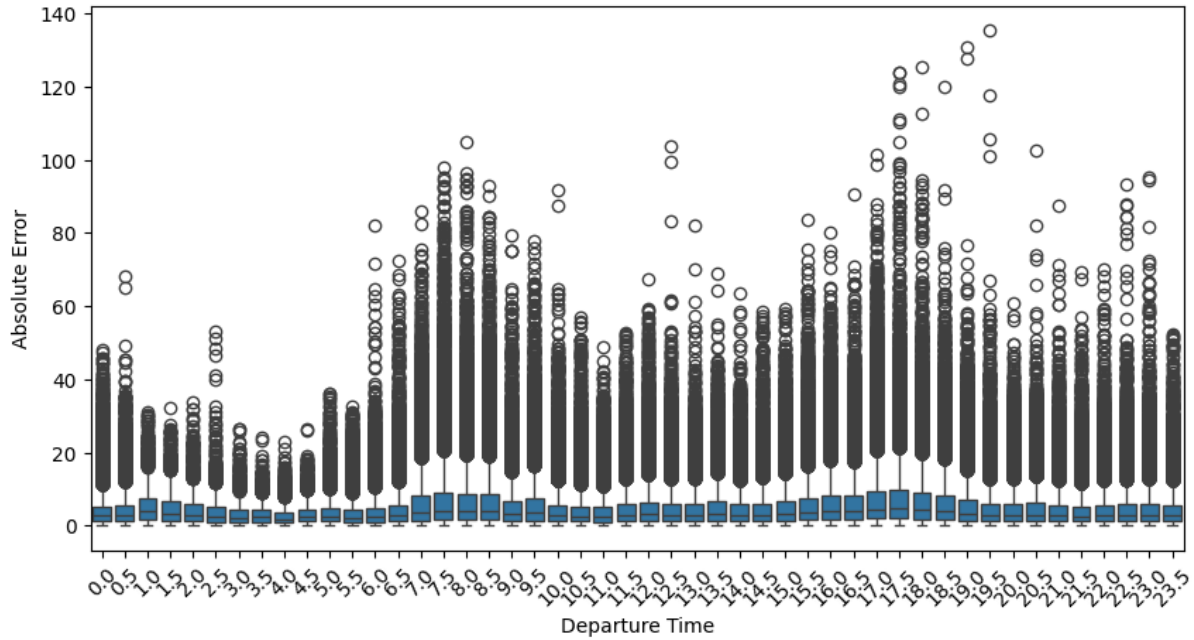
**Improvement**
**MAE: 11.8 -> 5.3**



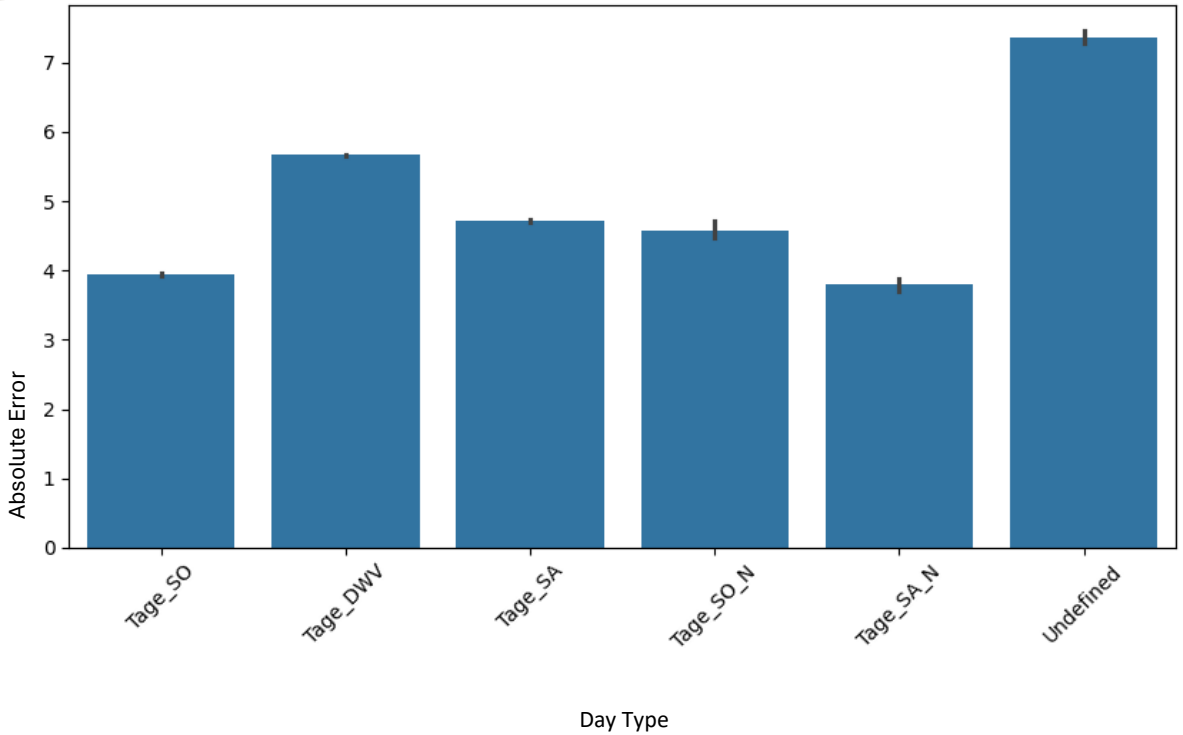



# Prediction: Deep Learning

## Box Plot of Absolute Error per Departure Time




## Absolute Error per Day Types




**Improvements:**

- Retaining information from previous inputs (timesteps), therefore, Recurrent Neural Networks (or other time-dependent model).

# Limitations

- Changing schedules, political decisions, societal behaviours and demographic shifts are especially unpredictable, making long-term precise predictions non-realistic, despite retrospective data analysis offer insight into the confidence of predictions.
- Datasets we used are limited on showing seasonalities, since the passenger numbers are averaged and we cannot guarantee the quality of the process of data collection.


Image from behance.net

# Thank you for your attention!!

Our website with result summary

Our GitHub repository