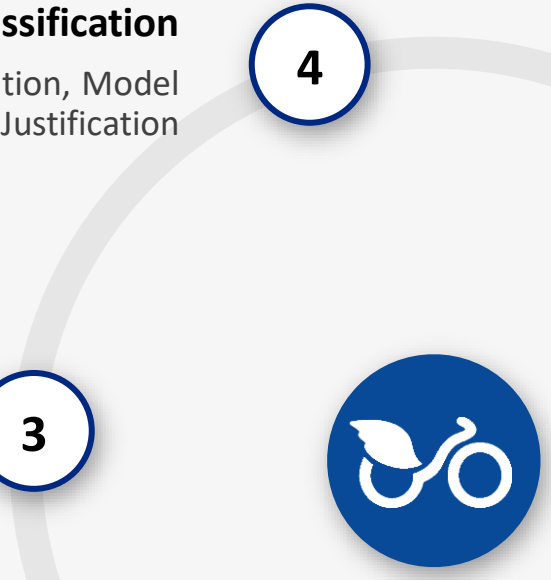
Project Report

PROGRAMMING DATA SCIENCE

University of Cologne – Manuel Gassner, Long Le, Irene(Jiaying) Xu

Project Design



1 Introduction

Project Timeline, Team Responsibilities, Project Overview

2 Data Processing

Initial Data Exploration, Data Cleaning, Data Manipulation

3 Data Analysis

Statistical Analysis, Data Visualization, etc.

4 Prediction & Classification

Predictors, Model Creation, Model Testing, Possible Justification

5 Conclusion & Discussion

Conclusion,
Suggestion for future analysis

Team Members and Responsibilities

Manuel Gassner

- Business understanding
- Data understanding
- Code refactoring
- QA
- Descriptive analysis of data
- Creation and evaluation of the prediction models
- Handling scm (GIT)

Irene Xu

- Business understanding
- Data understanding
- Code refactoring
- QA
- Descriptive analysis of data
- Creation of report and presentation slides
- Data description

Long Le

- Business understanding
- Data understanding
- Code refactoring
- QA
- Descriptive analysis of data
- Data Visualization
- Creation of report and presentation slides
- Creation and evaluation of the prediction models

Project Timeline


Business Understanding


Data Preparation


Data Visualization


Modelling


Validation & Evaluation


Report & Slides Creation


Test result & Presentation

Present -

Phase 1

Business Understanding – 2 days
Data Preparation / Understanding – 2 weeks
Data Visualization – 1 week

*Creating & maintaining the git repo

Manuel Gassner
Irene Xu
Long Le

23.05.2020 -

Phase 2

Model Creation – 1 week

*Code refactoring
*Maintaining the git repo

Manuel Gassner
Long Le

02.06.2020 -

Phase 3

Evaluation – 3 days
Report & Slides Creation – 3 days

*Cleaning up the git repo

Irene Xu
Long Le

08.06.2020 -

Phase 4

Getting test data
Presentation

Manuel Gassner
Irene Xu
Long Le

Introduction

nextbike

Nextbike is a German company that produces and operates bike sharing systems around the world.

- **Bike-sharing** is one of the **fastest growing** transportation services
- The ability to **predict bike-sharing patterns** allows service providers to be **more efficient** and **cost-effective**
- This project examines the 2019 **NextBike's trip data** of bike usage in **Frankfurt** and constructs prediction models of usage patterns.
- The goal is to explore **how the bike sharing systems are used**, and to gain insights into the **complex bike sharing activity patterns**.

Data Processing

The data set covers approximately 532,000 entries of the year 2019. There are 13 columns in total, the types and meanings of the columns are stated in following Table.

Variable Name	Format	Description
b_number	int64	ID of the bike being rented
b_bike_type	float	Types of the bike, there are 4 types of bike
p_spot	bool	If the bike is in the bike station, T = in station, F = not in station
p_bike	bool	If the bike is floating, T = in, F = not floating
p_name	object	Human readable name of the bike station
p_number	float	ID of the bike station
p_place_type	int64	Types of places, e.g., 7 = hotels, 3 = education institutions
p_uid	int64	Unique ID of the location
p_bikes	int64	The number of available bikes in the station
p_lat	float	latitude of the start/end location
p_lng	float	longitude of the start/end location
datetime	object	YY-MM-DD 00:00:00, date-time of the booking
trip	object	Status of the booking (start/end, first/last)

Table 1: Variables of NextBike's data set

Data Cleaning

"First" and "last" data that only represent the availability of the bike but not the actual trip data.

Rides that have more than one "start" or "end" data, where the bike is already occupied and a start booking appears when the later booking is dropped

Rides ≤ 3 minutes with no change in location. An actual ride does not take place.

Rides ≥ 2 hours, as 93% of trip is between 2 - 120 minutes, such rides are defined as outliers.

Rides that show negative/wrong coordinates. This could happen if an error occurs at the bike station/spot while returning a bike.

Other outliers, e.g., batch bookings that caused by bike reallocation

*Data objects that show the above characteristics are removed from the original data set

New DataFrame

	Bike_number	Start_Time	End_time	Start_Longitude	Start_Latitude	End_Longitude	End_Latitude	Start_Station	End_Station_number	Duration	Bikes_on_position	End_Bikes	Zip_codes	month
13	11451	2019-05-30 15:42:00	2019-05-30 17:22:00	8.553437	50.101748	8.554450	50.101711	4291.0	0.0	100.0	5	1	65929	05
14	11451	2019-05-31 02:22:00	2019-05-31 02:32:00	8.554450	50.101711	8.556589	50.114979	0.0	0.0	10.0	1	1	65929	05
22	11542	2019-09-10 06:53:00	2019-09-10 07:28:00	8.664168	50.107599	8.611191	50.124972	4240.0	0.0	35.0	9	1	60329	09
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Descriptive Analysis - Outliers

Rides ≥ 2 hours, as 93% of trip is between 2 - 120 minutes, such rides are defined as outliers.

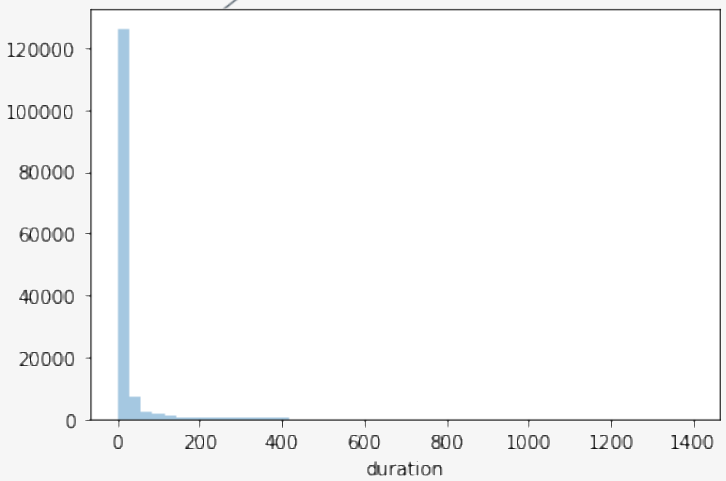


Figure 1: Distribution (with outliers)

*Range from 2 minutes to 1394 minutes

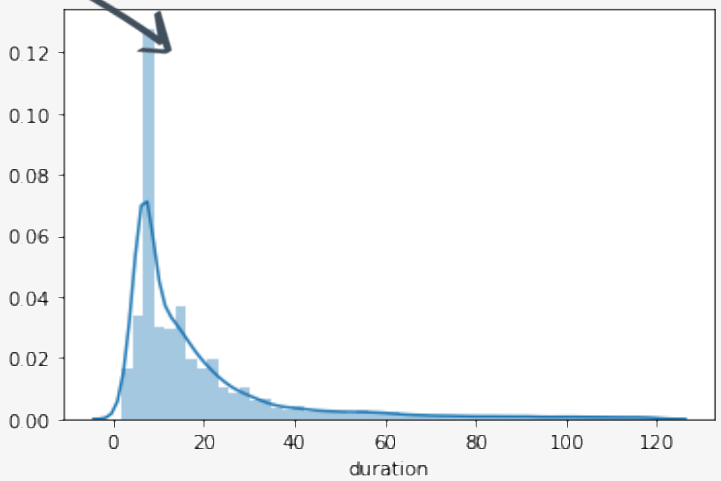


Figure 2: Distribution (without outliers that ≥ 2hrs)

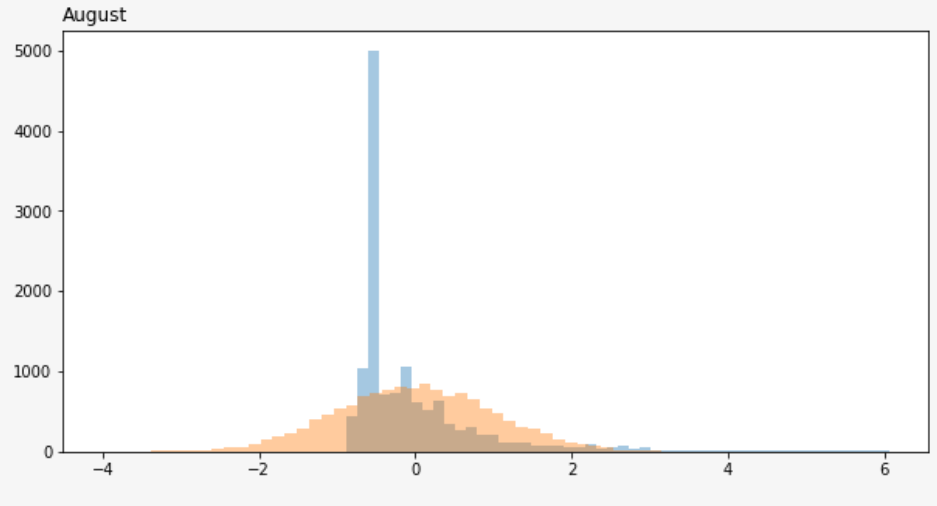


Figure 3: Distribution for August (without outliers that ≥ 2hrs)

Descriptive Analysis - Outliers

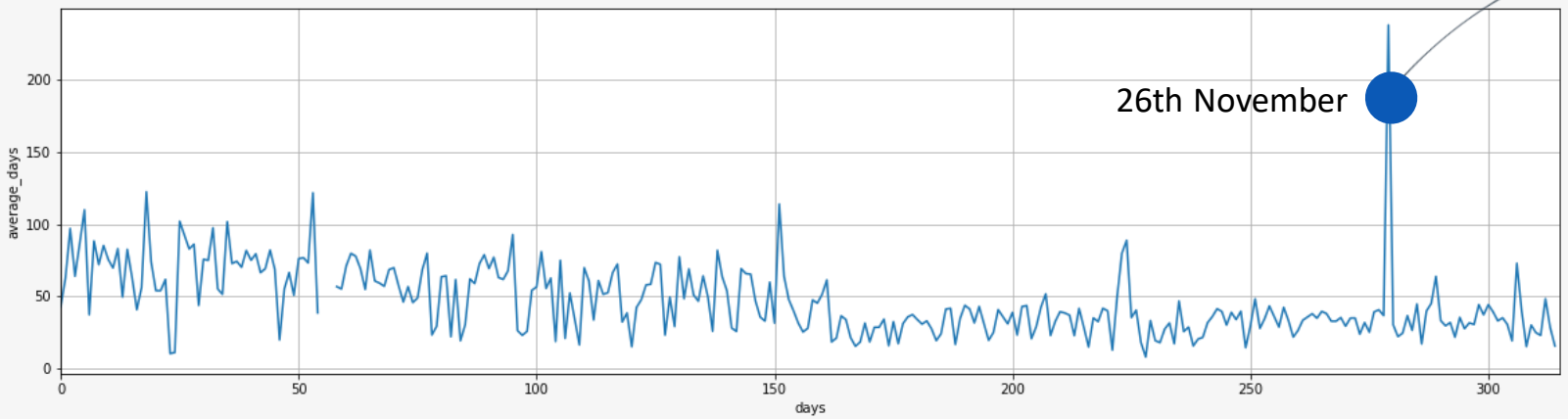


Figure 4: Weekly average duration by days

A "batch" booking of 250 bikes for 384 minutes on the same time at 1 a.m.
Assumption: Reallocation of bikes to the station by Nextbike's employees

Remaining trip data: 51,400 Entries / Trips

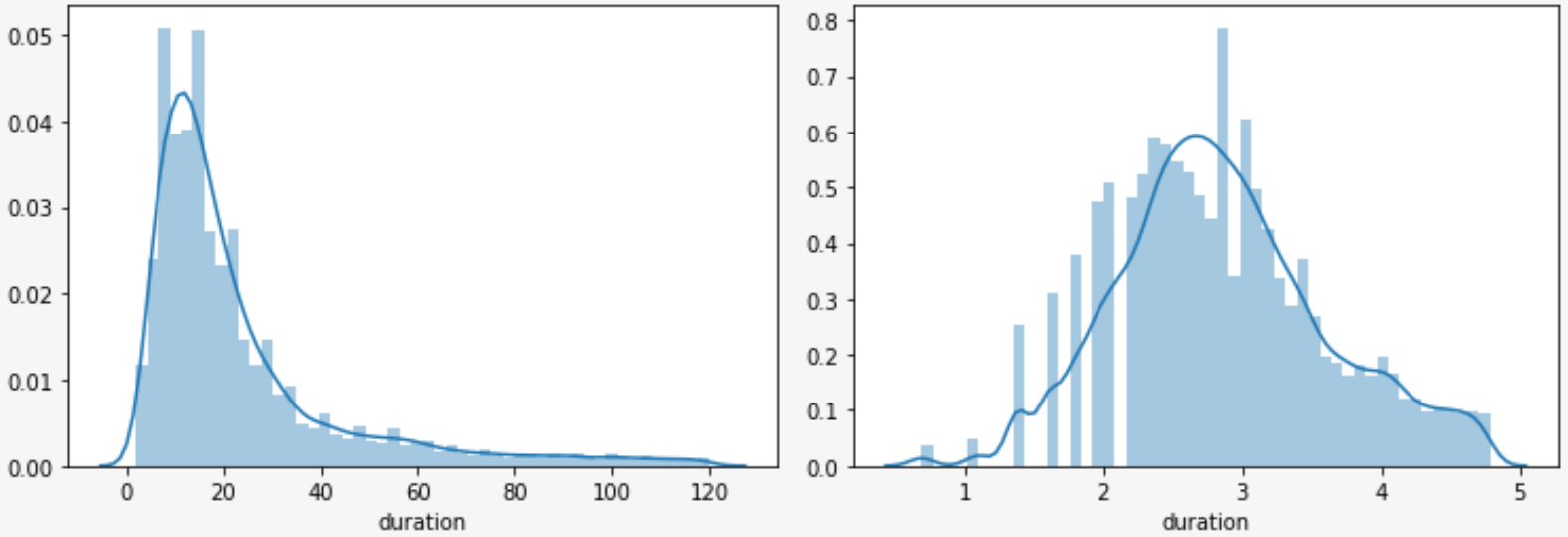


Figure 5: The distribution of duration <120 before log transformation

Figure 6: The distribution of duration <120 after log transformation

After dropping all the outliers that would affect our predictions, the overall distribution shape for the remaining 51400 trips goes close to the shape of the normal distribution.

We will use the data set with a log transformation of the target value for our further analysis prediction.

Descriptive Analysis – Aggregate Statistics

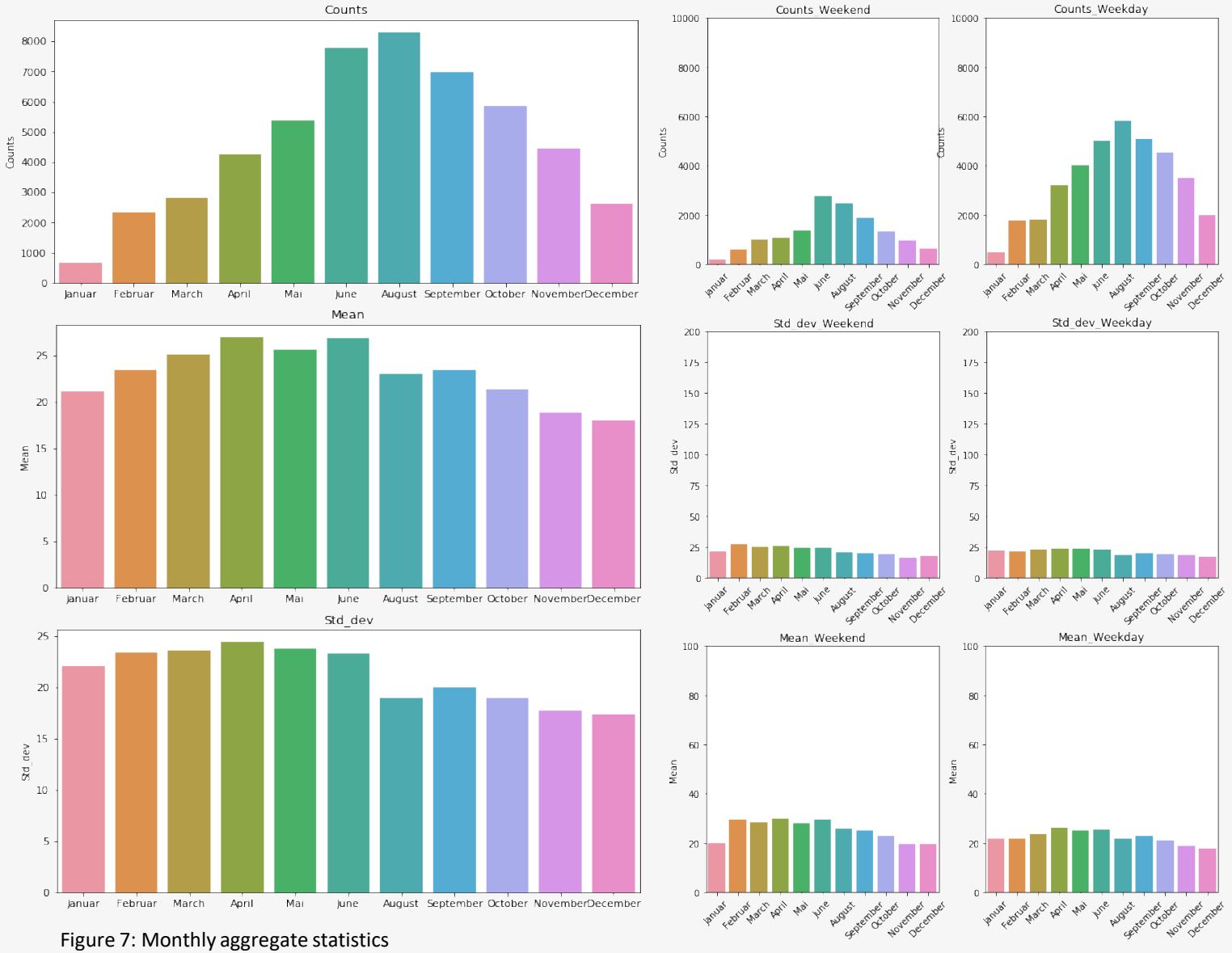


Figure 7: Monthly aggregate statistics

Seasonal Trend

More trips in summer month (June, July, August) than winter month (Dec, Jan, Feb).

Lower demand for bike sharing services in winter month.

Weekends vs. Weekdays

(Except winter month), longer trip duration on weekends; shorter trip duration on weekdays.

More trips occurred on weekdays ; less trips occurred on weekends.

Possible travel patterns:

In general, the weekday bike travel is more in favor compared to weekend bike travel when there's a commuting need.

Descriptive Analysis – Aggregate Statistics

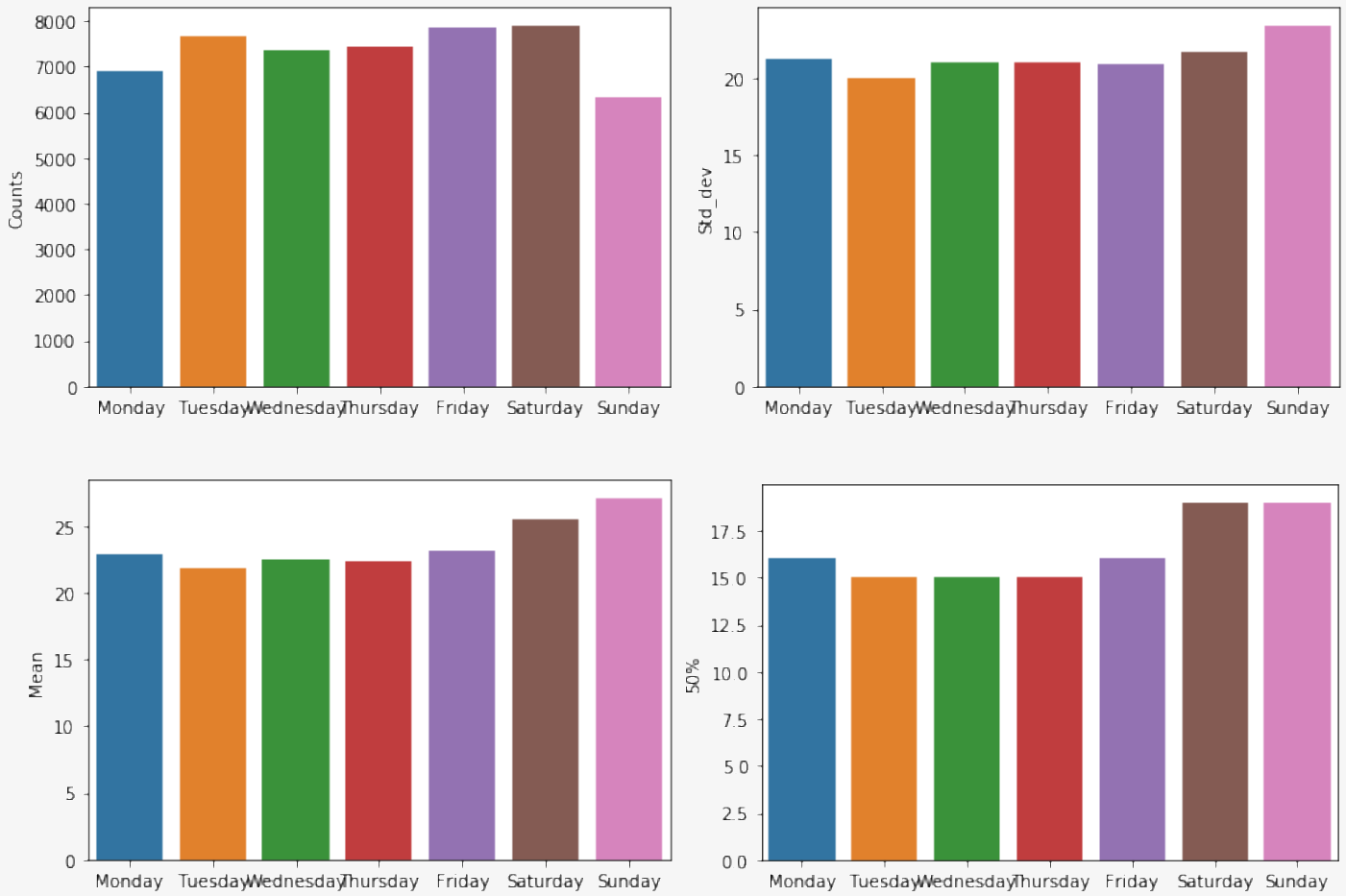


Figure 8: Weekly aggregate statistics

Weekends vs. Weekdays

Longer trip duration on weekends; shorter trip duration on weekdays.

More trips occurred on weekdays (+Saturday); less trips occurred on Sunday.

Possible travel patterns:

Weekend travel behavior is different from the weekday travel behavior.

Descriptive Analysis – Aggregate Statistics

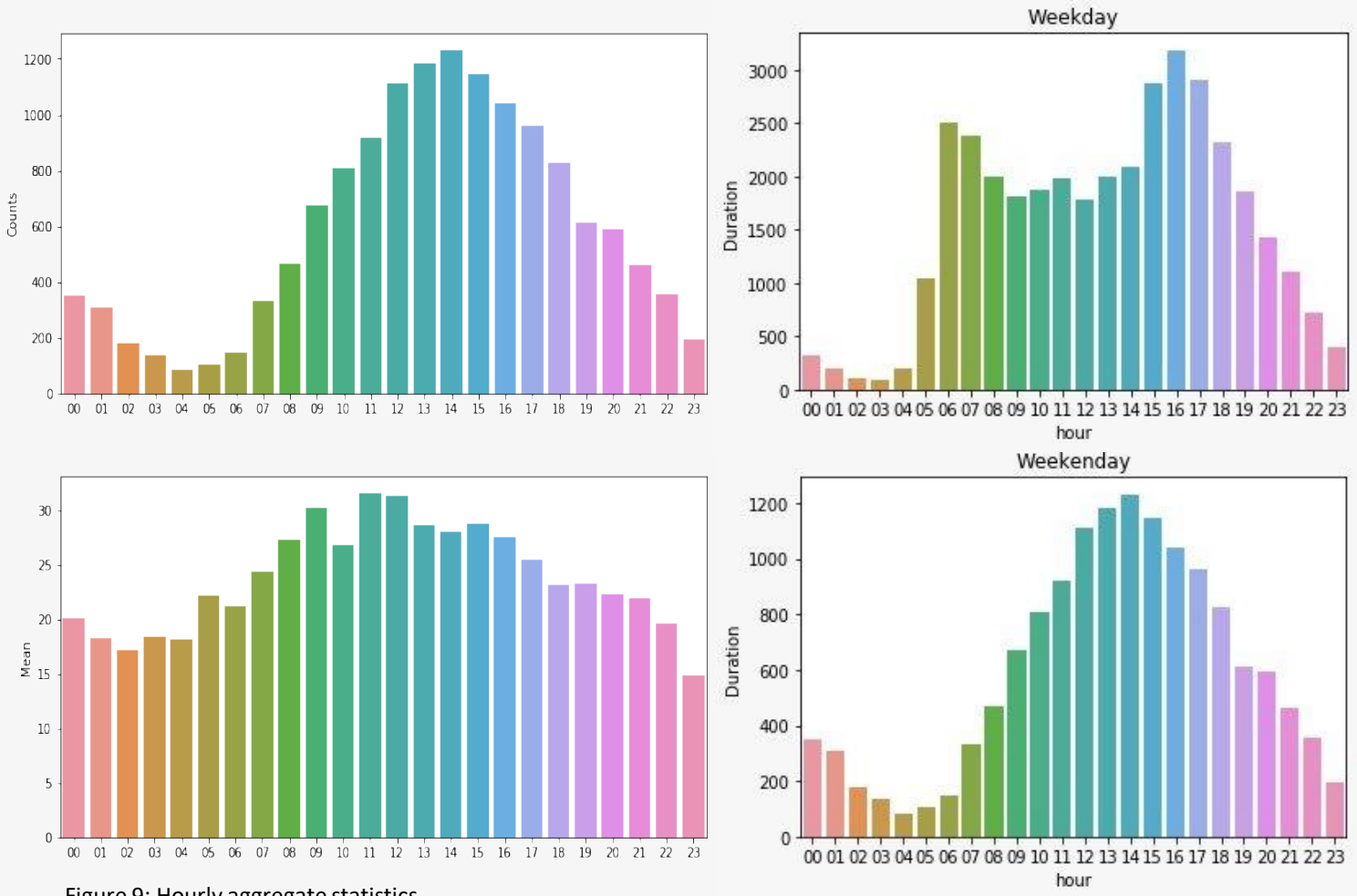


Figure 9: Hourly aggregate statistics

Hourly Difference

The average duration reaches its peak in the morning commuting hours between 8-10 o'clock (go to work or go to school) and 12 o'clock (lunch hour), whereas the trough is from midnight to early morning.

Weekends vs. Weekdays

People are more likely to use the shared bike in the afternoon (12-3pm) on weekends, and in the commuting hours (6am-8am) on weekdays.

Possible travel patterns:

Weekday bike travel is more in favor compared to weekend bike travel in commuting hours

Data Visualization

Geographical distribution of starting locations of the trips in summer months

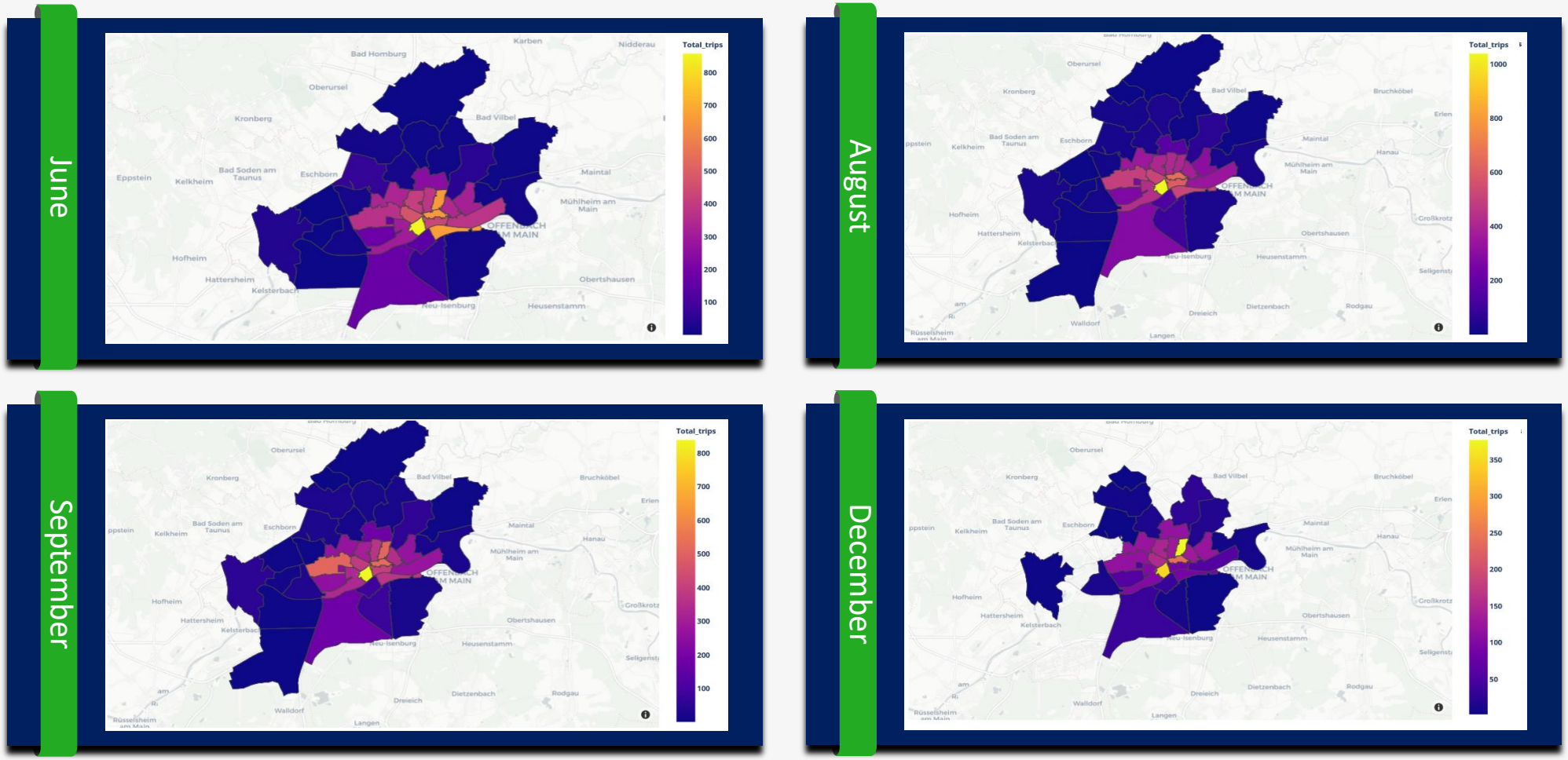


Figure 10: Map of started trips by Postal Region in June, August, September, December, 2019

Pattern

More trips are generated within the city center than the surrounding areas.
In other words, the closer the region is to the city center, the more trips are generated in that region.

Data Visualization

Average duration of the trips by the starting PLZ regions

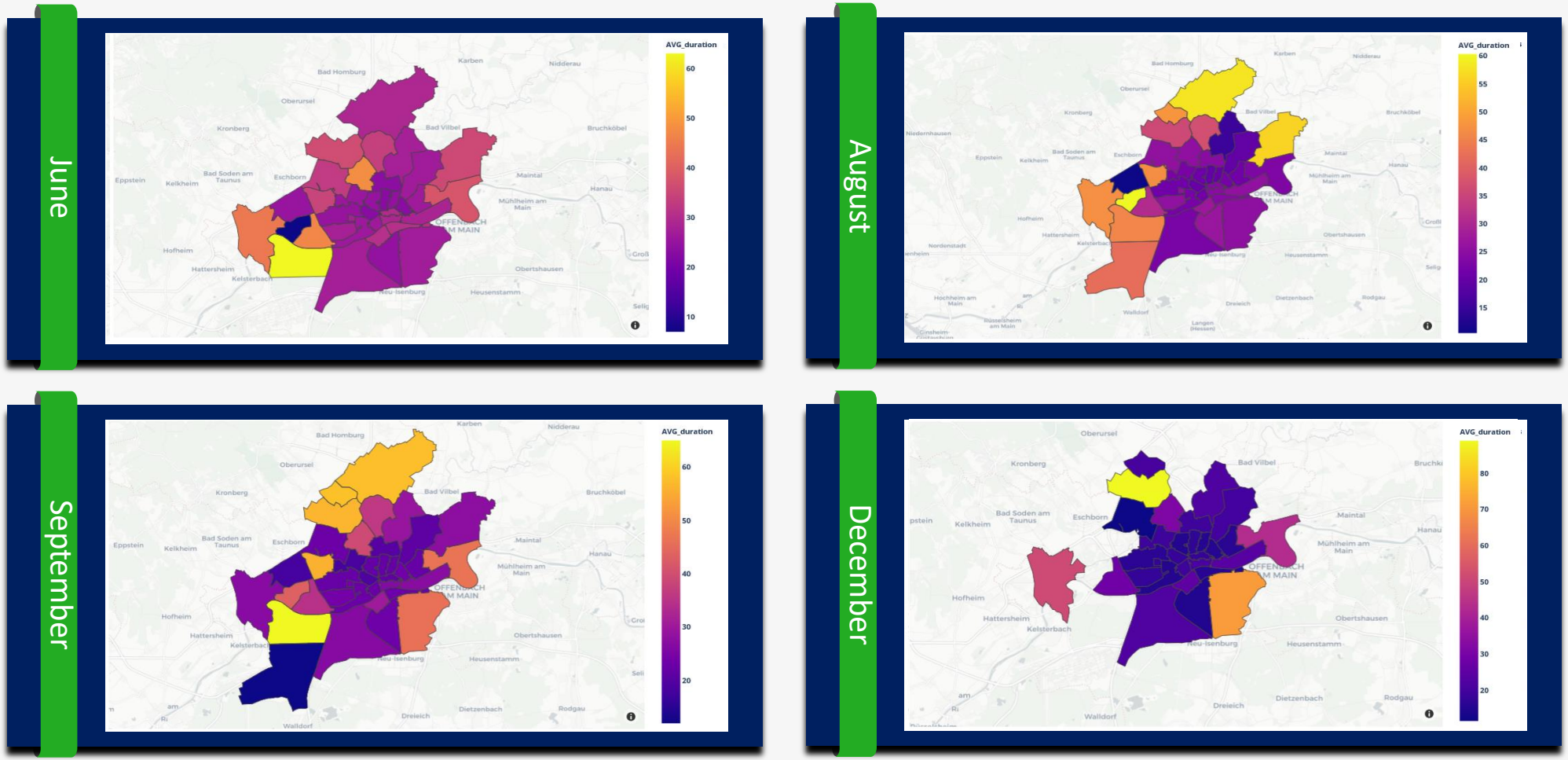


Figure 11: Map of average trip duration by started Postal Region in June, August, September, December, 2019

Pattern

The average duration of trips that started in the center regions were shorter than that of trips started in the border regions. We hypothesized that the region from which the trips started (Border_district) could be a predictor for trip duration.

Data Visualization

Explore whether people would use bikes to travel to public events.

IAA
» Driving tomorrow

Location: Messe Frankfurt exhibition center
Visit day: From 12 to 22 September, 2019
Visit hours: Daily from 09:00 am to 07:00 pm, 20th of September from 11 am to 9 pm

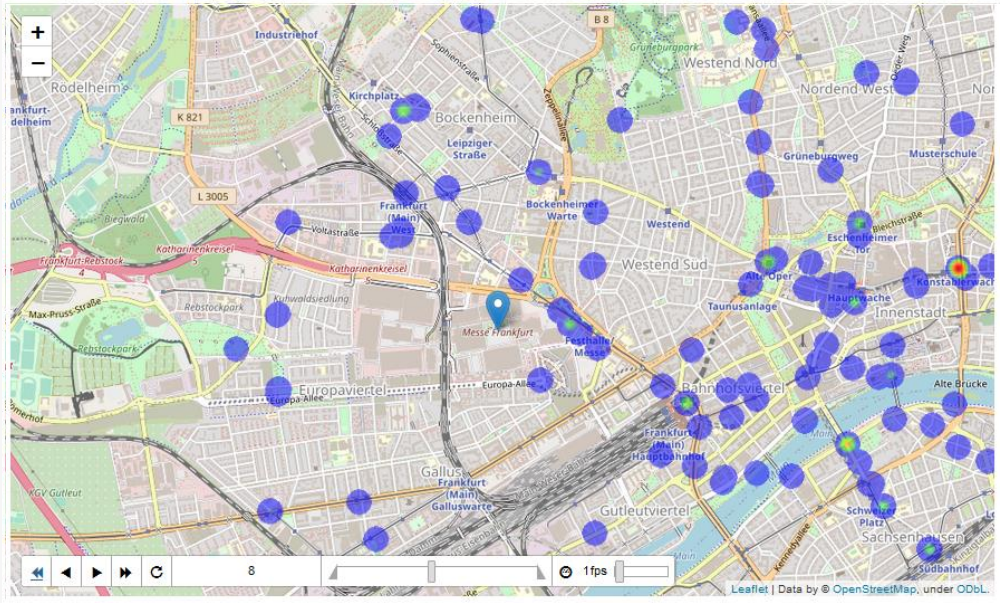


Figure 12: Heatmap of the end trips near the exhibition center on Saturday, 14 September 2019.

Football match, Eintracht Frankfurt vs. Dortmund
Location: Commerzbank-Arena, Frankfurt
Date and time: 18:00 CEST on 22 September, 2019
Attendance: 51,500

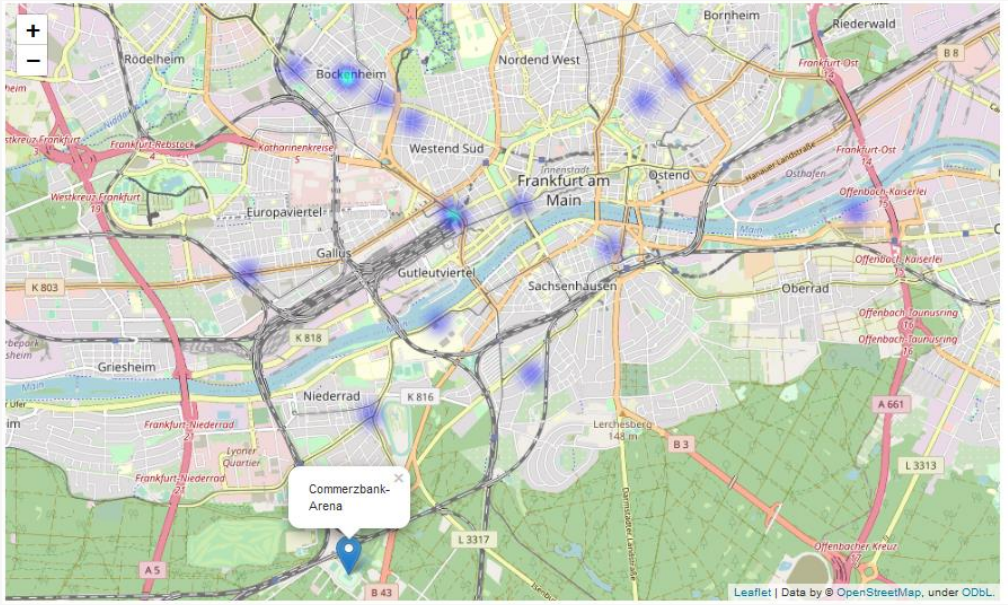


Figure 13: Heatmap of the end trips 1.5 hours before the football match between Eintracht Frankfurt and Dortmund started on 22 September, 2019

Data Visualization

Explore whether people would use bikes to travel to public events.

IAA
» Driving tomorrow

Location: Messe Frankfurt exhibition center
Visit day: From 12 to 22 September, 2019
Visit hours: Daily from 09:00 am to 07:00 pm, 20th of September from 11 am to 9 pm

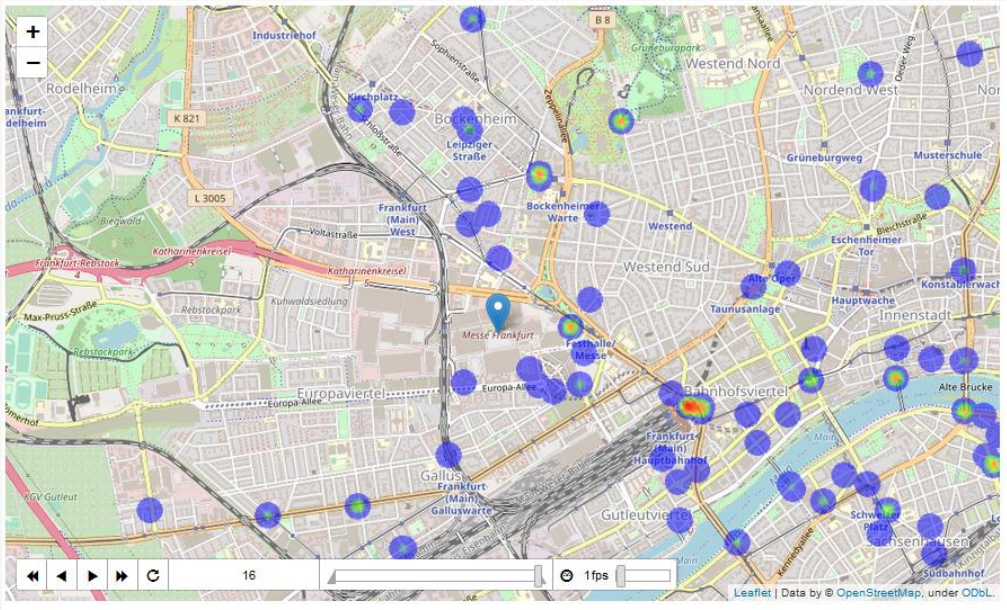


Figure 14: Heatmap of the end trips near the exhibition center on the last day of the event.

Football match, Eintracht Frankfurt vs. Dortmund
Location: Commerzbank-Arena, Frankfurt
Date and time: 18:00 CEST on 22 September, 2019
Attendance: 51,500

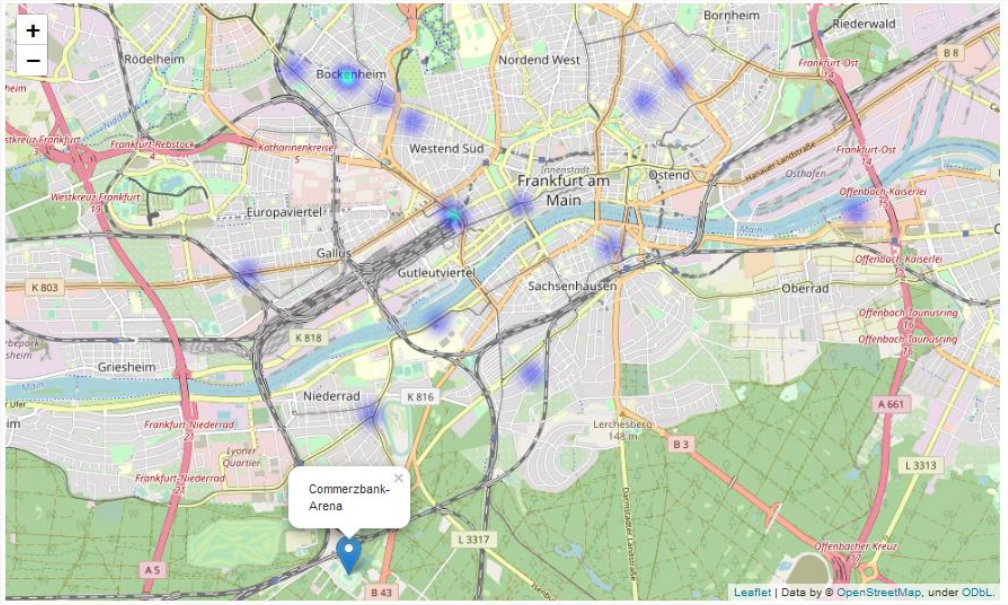


Figure 15: Heatmap of the end trips 1.5 hours before the football match between Eintracht Frankfurt and Dortmund started on 22 September, 2019

Data Visualization

Explore whether people would use bikes to travel to public events.

IAA

» Driving tomorrow

Location: Messe Frankfurt exhibition center
Visit day: From 12 to 22 September, 2019
Visit hours: Daily from 09:00 am to 07:00 pm, 20th of September from 11 am to 9 pm

- There were trips that ended next to the exhibition center during the opening hours.
- Weekends captured more trips to the event location than that of normal.
- As the event came closer to its end, more trips that ended near the exhibition center were captured in the map.
- Interestingly, 2 days before the first opening day, there were trips that ended inside the exhibition center. This could be explained that event staff and volunteers might have organized their preparation meeting before the event started.

Figure 14: Heatmap of the end trips near the exhibition center on the last day of the event.

Football match, Eintracht Frankfurt vs. Dortmund
Location: Commerzbank-Arena, Frankfurt
Date and time: 18:00 CEST on 22 September, 2019
Attendance: 51,500

- To investigate if people commuted to the football match using bikes, data for the trips that ended 1.5 hour before the match started was employed in the visualization.
- In contrast to the IAA event, the map shows nearly no bike trips ended near the Commerzbank-Arena before the match started.
- The location of the Commerzbank-Arena might be a decisive factor in this case, as the stadium was far from the city center, the bike route to commute from the city center to the stadium might be difficult.

Figure 15: Heatmap of the end trips 1.5 hours before the football match between Eintracht Frankfurt and Dortmund started on 22 September, 2019

Data Visualization

Explore if students used bikes to travel to university.

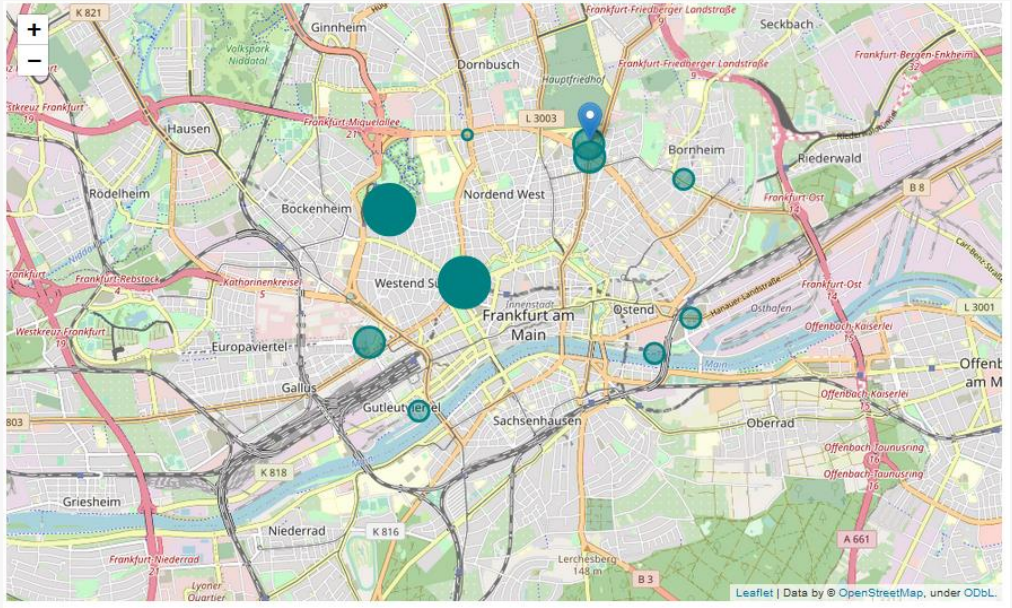


Figure 16: Map of the number of bikes at fixed station from 7AM to 8AM on May 03, 2019.

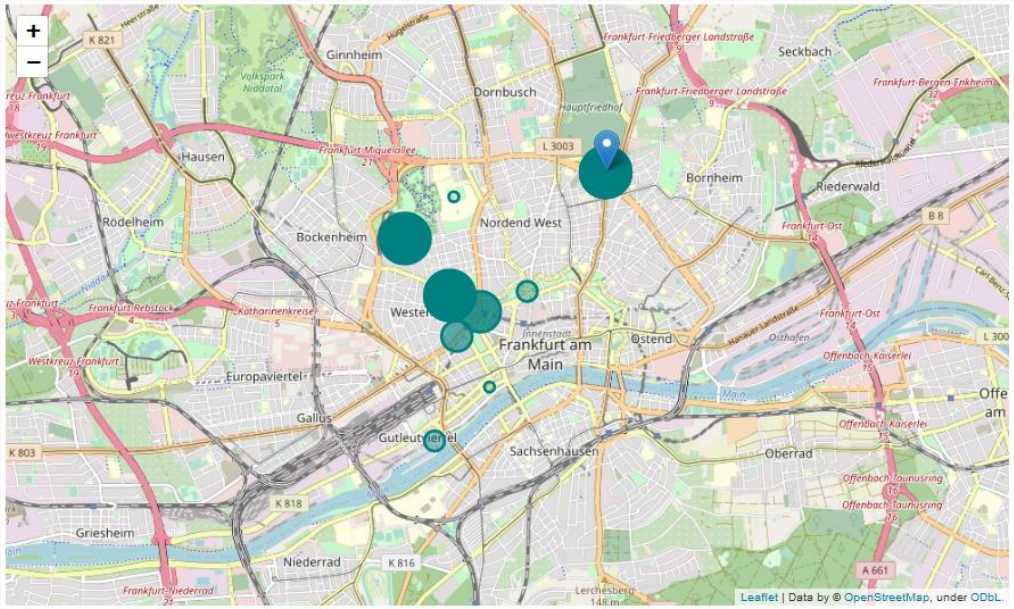


Figure 17: Map of the number of bikes at fixed station from 7AM to 8AM on May 08, 2019.

Pattern

From 7AM to 8AM (before normal lecture time) on the selected days (May 03 and May 08th 2019, normal weekday), the stations near the University of Applied Sciences were among the top stations with more bikes.

Data Visualization

Explore if students used bikes to travel to university, monthly analysis

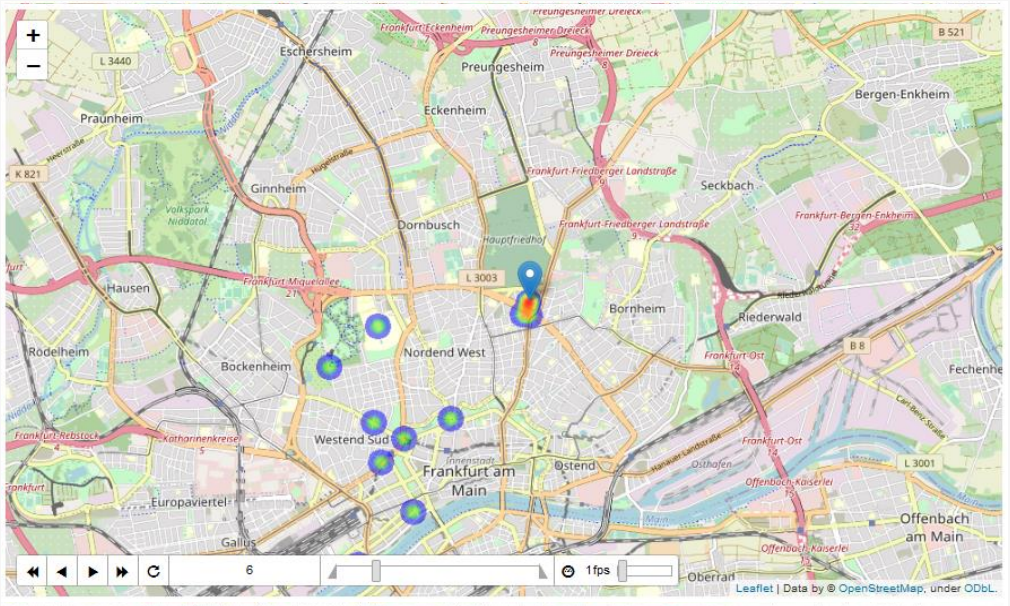


Figure 18: Heatmap of the location of the trips that ended from 7AM to 8AM on May 08, 2019.

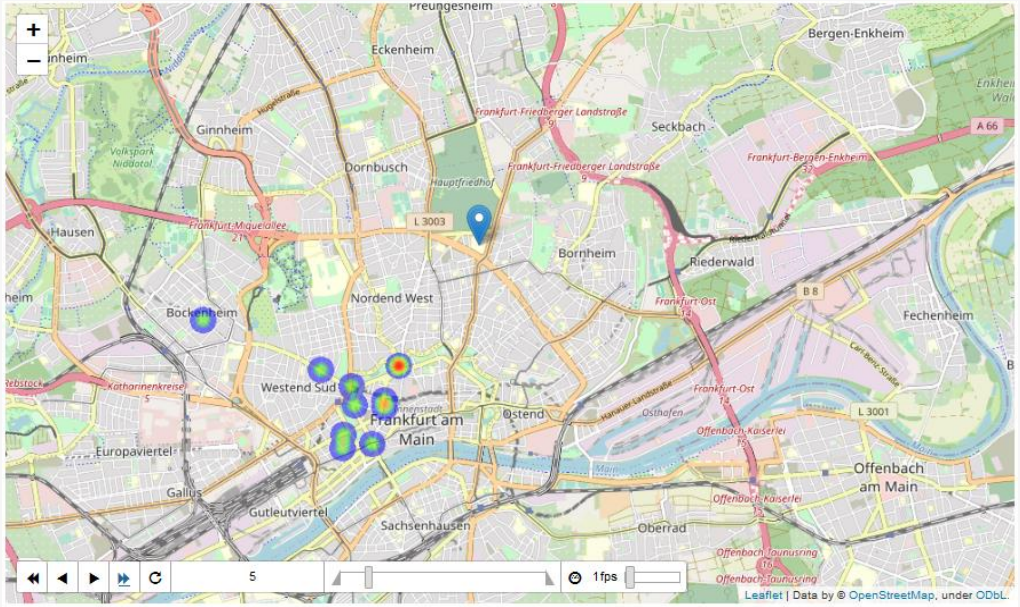


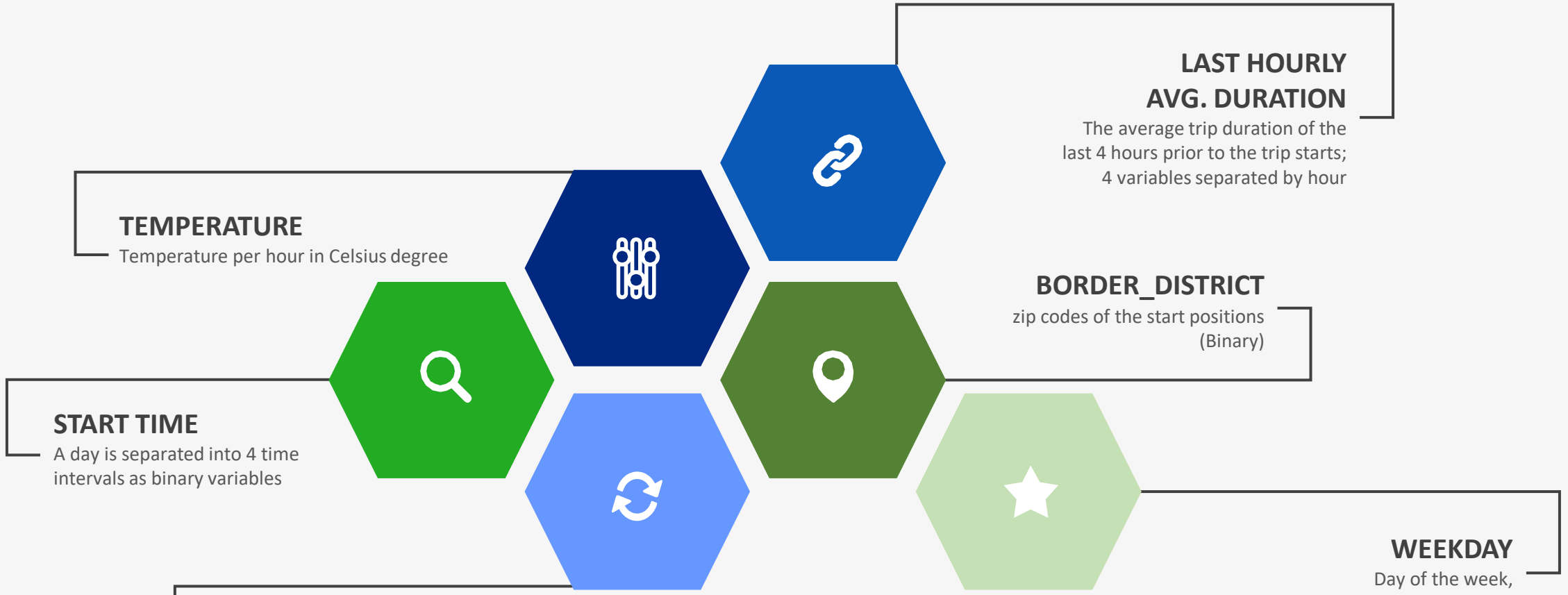
Figure 19: Heatmap of the location of the trips that ended from 7AM to 8AM on August 05, 2019.

Pattern

The areas near the University of Applied Sciences were the top popular locations for bike trips to end from 7AM to 8AM in most of the days in May, September, and October (except for weekends and public holidays). In contrast, August has very few trips ended near the university location from 7AM to 8AM.

Such pattern could be explained by the fact that August is the summer vacation for university students. When the spring semester ends on August 1st, there were many trips that ended near the university captured. But after that, very few trips ended within the selected time frame.

Predictors – for prediction model to predict trip duration



LAST HOURLY AVG. DURATION
The average trip duration of the last 4 hours prior to the trip starts; 4 variables separated by hour

TEMPERATURE
Temperature per hour in Celsius degree

BORDER_DISTRICT
zip codes of the start positions (Binary)

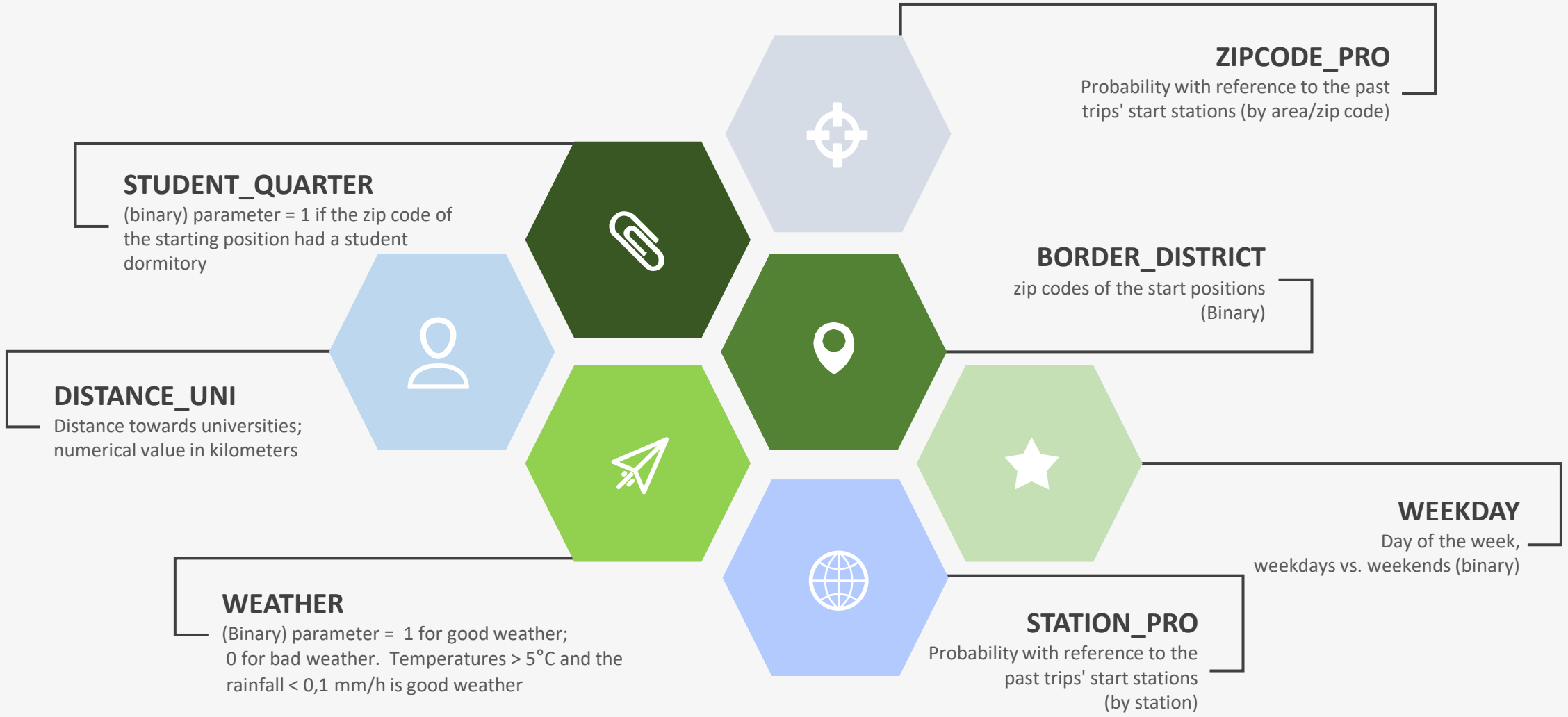
START TIME
A day is separated into 4 time intervals as binary variables

WEEKDAY
Day of the week, weekdays vs. weekends (binary)

LAST DAILY AVG. DURATION
Average duration of the last 2 days prior to the trip starts; 2 variables separated by day

Predictors - for classification model

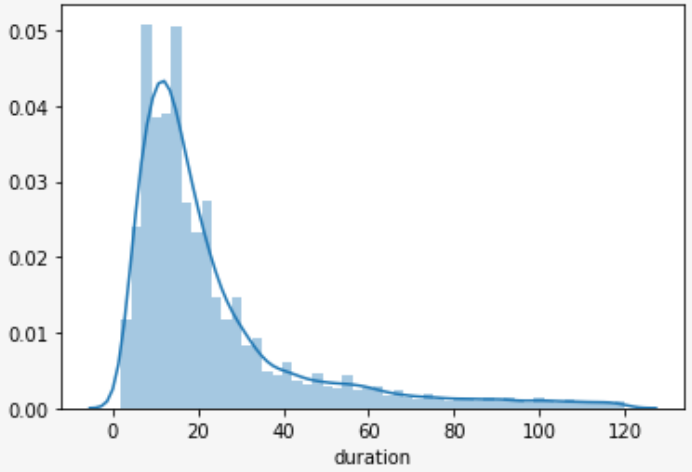
to classify if a trip is travelling towards the university



Prediction Methods

Prediction Model

… to predict trip duration



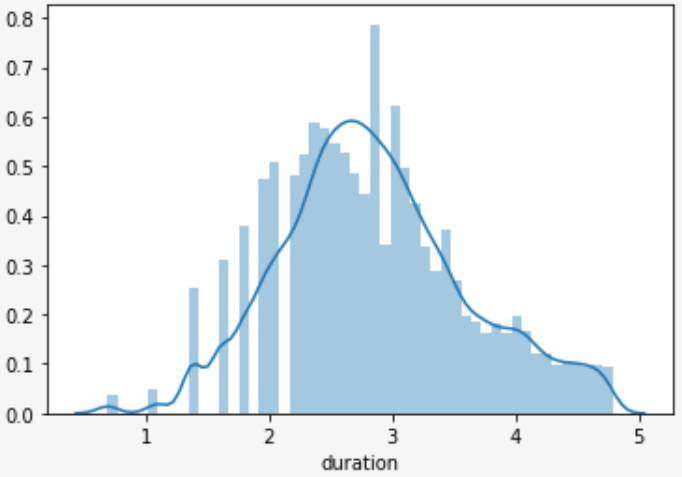
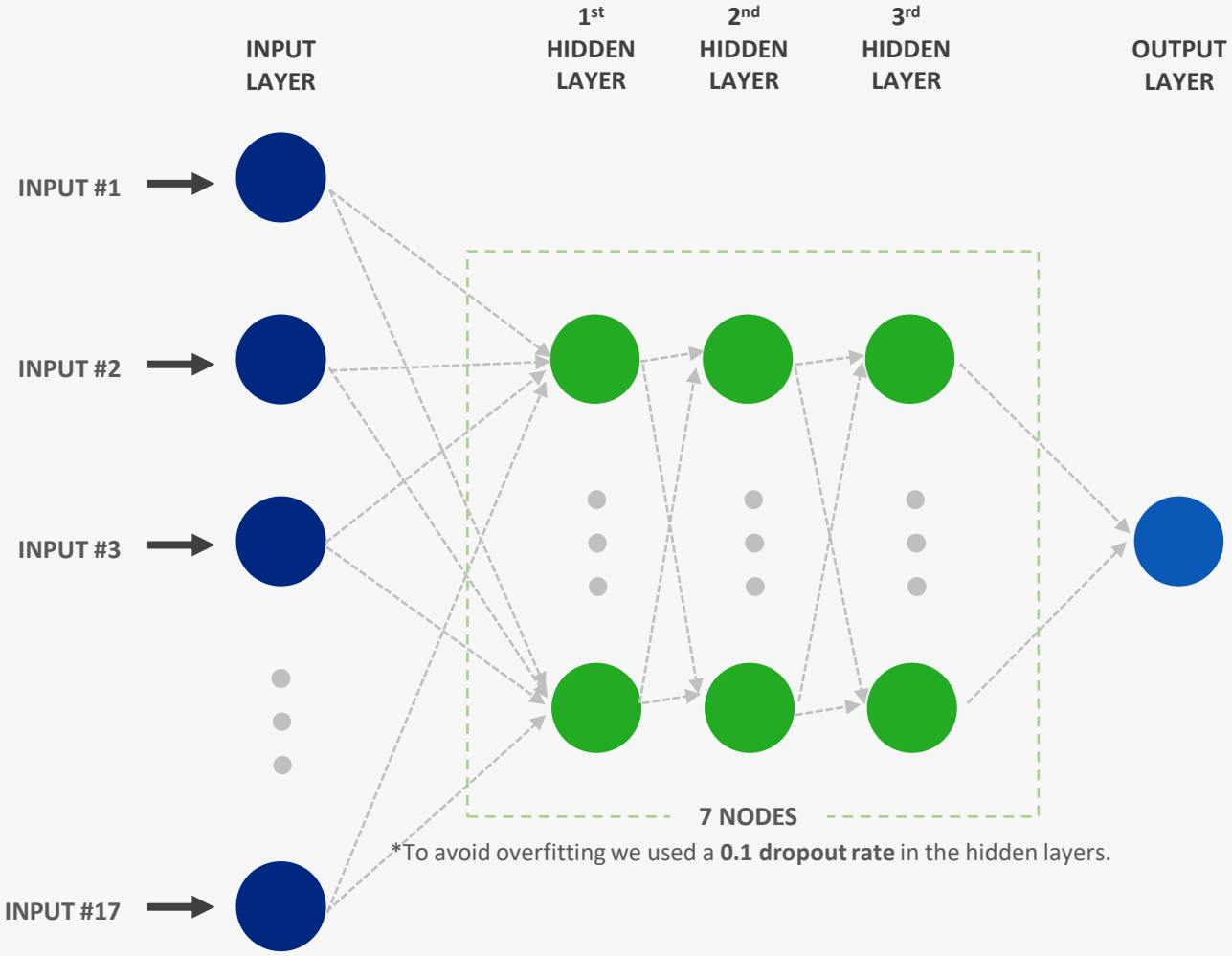


Figure 5: The distribution of duration <120 before log transformation

Figure 6: The distribution of duration <120 after log transformation

Classification Model

… to classify if a trip is travelling towards the university



*To avoid overfitting we used a **0.1 dropout rate** in the hidden layers.

Evaluation for Prediction Model

Metrics used to evaluate models performance:

- **Mean-absolute-error (MAE):** general miss on prediction in minutes
- **Root-mean-square-error (RMSE):** used to detect outliers
- **R^2,** indicator for the degree of generalizability of the model.

Metrics	Validation Set	Test Set
MAE	13 mins	15.7 mins
RMSE	21 mins	15.9 mins
R^2	0.051	0.008

Table 2: The metrics results for model evaluation (prediction model)

Findings

- Underfitted model
- Prediction on average far of real values
- Reengineering required for model and predictors
- Model did not fit the underlying data structure

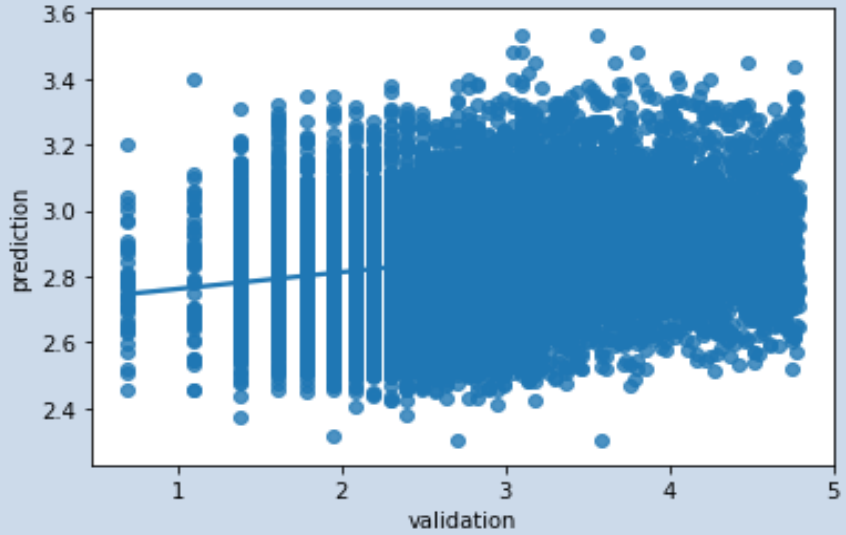


Figure 20: performance of prediction model (validation set)

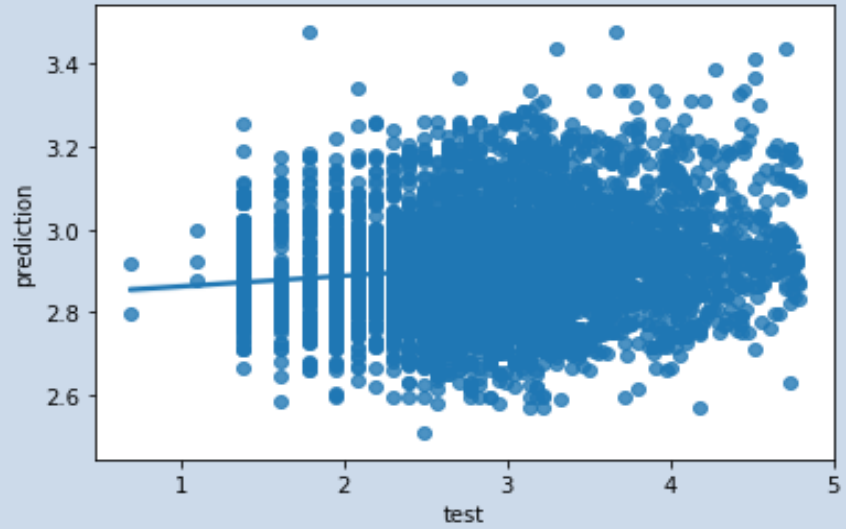


Figure 21: performance of prediction model (test set)

Evaluation for Classification Model

Metrics used to evaluate model performance:

- **Confusion matrix** shows us how good the model classifies the data

Confusion matrix	Validation Set	Test Set
True/Positive	5208	2326
True/Negative	3511	1166
False/Positive	3993	917
False/Negative	2706	988
Classification Accuracy	0.52% ~ 52%	0.647% ~ 65%

Table 3: The confusion matrix results for model evaluation (classification model)

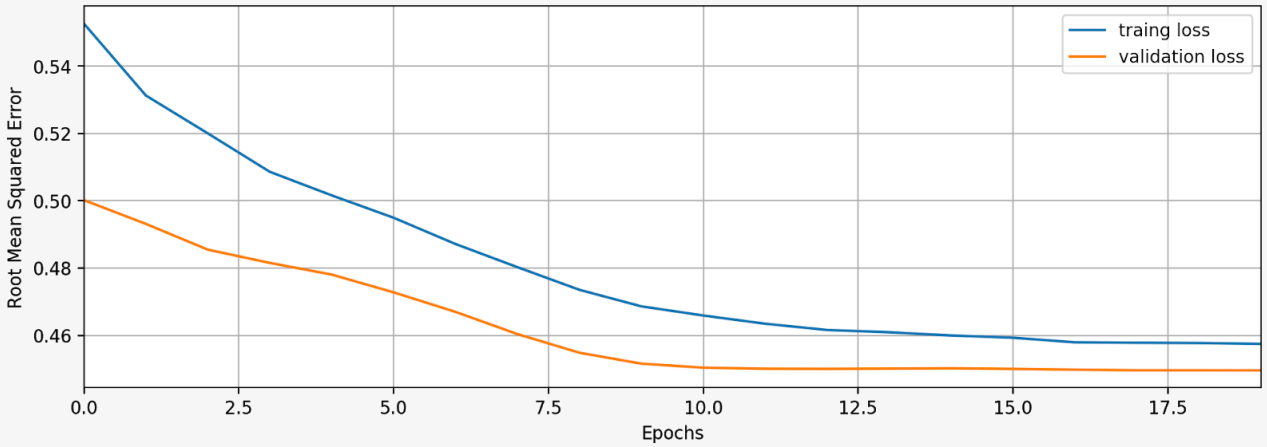


Figure 22: The loss and validation lost of the models with their number of epochs

Findings

- Better performance on test set
- Results not clear
- Small test set
- Blackbox mentality of neuronal networks

Conclusion

Conclusion from Descriptive Analysis

1. Higher demand for bike sharing services in summer month
2. Longer rental period for bike sharing services on weekends.
3. In general, the weekday bike travel is more in favor compared to weekend bike travel when there's a commuting need.

Conclusion from Data Visualization

1. Higher demand, but shorter rental period for bike sharing services in city center.
2. University students can be the target users of bike-sharing services.
3. Public events have impact on NextBike's bike-sharing usage.

Conclusion from Model Development

1. The data set contains a large amount of outliers, noises, and measurement errors
2. Unable to find a good indicator for predicting trip duration, or direction towards the universities.
    1. Only in the summer month the temperature seems to show a correlation to the duration of the trips.
3. The validity of the prediction and classification models fit are questionable, the identification of outliers is crucial for further investigation.

Reference

McLeod, K. (2014). New: Census Data on Bike Commuting. [online] League of American Bicyclists. Available at: https://bikeleague.org/content/new-census-data-bike-commuting

Nextbike GmbH (2020). Frequently Asked Questions. [online] www.nextbike.de. Available at: https://www.nextbike.de/en/faq/.